International Journal on Advances in Intelligent Systems

2025 vol. 18 nr. 3&4

The *International Journal on Advances in Intelligent Systems* is Published by IARIA.
ISSN: 1942-2679
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Responsibility for the contents rests upon the authors and not upon IARIA, nor on IARIA volunteers, staff, or contractors.

IARIA is the owner of the publication and of editorial aspects. IARIA reserves the right to update the content for quality improvements.

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy or print, providing the reference is mentioned and that the resulting material is made available at no cost.

Reference should mention:

International Journal on Advances in Intelligent Systems, issn 1942-2679
vol. 18, no. 3&4, year 2025, http://www.iariajournals.org/intelligent_systems/

The copyright for each included paper belongs to the authors. Republishing of same material, by authors or persons or organizations, is not allowed. Reprint rights can be granted by IARIA or by the authors, and must include proper reference.

Reference to an article in the journal is as follows:

<Author list>, "<Article title>"
International Journal on Advances in Intelligent Systems, issn 1942-2679
vol. 18, no. 3&4, year 2025, <start page>:<end page> , http://www.iariajournals.org/intelligent_systems/

IARIA journals are made available for free, proving the appropriate references are made when their content is used.

Sponsored by IARIA
www.iaria.org

Copyright © 2025 IARIA

International Journal on Advances in Intelligent Systems
Volume 18, Number 3&4, 2025

Editors-in-Chief

Hans-Werner Sehring, NORDAKADEMIE - University of Applied Sciences, Germany
Steve Chan, Decision Engineering Analysis Laboratory, USA

Editorial Board

Siby Abraham, NMIMIS Deemed to be University, Mumbai, India
Abdulelah Alwabel, Prince Sattam Bin Abdulaziz University, Saudi Arabia
Razvan Andonie, Central Washington University, USA
Andreas Aßmuth, Technical University of Applied Sciences OTH Amberg-Weiden, Germany
Isabel Azevedo, ISEP-IPP, Portugal
Mark Balas, Texas A&M University, College Station, USA
Leila Ben Ayed, National School of Computer Science | University la Manouba, Tunisia
Freimut Bodendorf, University of Erlangen-Nuremberg, Germany
Karsten Böhm, FH Kufstein Tirol - University of Applied Sciences, Austria
Ozgu Can, Ege University, Turkiye
Massimiliano Caramia, University of Rome Tor Vergata, Italy
Coskun Cetinkaya, Kennesaw State University, USA
Steve Chan, Decision Engineering Analysis Laboratory, USA
Mirela Danubianu, "Stefan cel Mare " University of Suceava, Romania
Leandro Dias da Silva, Universidade Federal de Alagoas, Brazil
Roland Dodd, CQUniversity, Australia
Larbi Esmahi, Athabasca University, Canada
Leila Ghomari, Ecole Nationale Supérieure des Technologies Avancées - ENSTA, Algiers, Algeria
Todorka Glushkova, University of Plovdiv, Bulgaria
Victor Govindaswamy, Concordia University Chicago, USA
Gregor Grambow, Aalen University, Germany
Maki K. Habib, The American University in Egypt, New Cairo, Egypt
Yousif A. Hamad, Siberian Federal University, Russia / Imam Ja'afar Al-Sadiq University, Iraq
Ridewaan Hanslo, University of Pretoria, South Africa
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Ahmed Kamel, Concordia College, Moorhead, USA
Rajkumar Kannan, Bishop Heber College (Autonomous), India
Leoneed M. Kirilov, Institute of Information and Communication Technologies - Bulgarian Academy of Sciences, Bulgaria
Dalia Kriksciuniene, Vilnius University, Lithuania
Panos Linos, Butler University, USA
Sandra Lovrencic, University of Zagreb, Croatia
Marcio Mendonça, Federal University of Technology - Paraná, Brazil
Harald Milchrahm, Technical University Graz | Institute for Software Technology, Austria
Shahab Mokarizadeh, Royal Institute of Technology (KTH), Stockholm, Sweden
Fernando Moreira, REMIT | Universidade Portucalense, Portugal
Joan Navarro, La Salle Campus Barcelona - Universitat Ramon Llull, Spain
Filippo Neri, University of Naples "Federico II", Italy
Rui Neves Madeira, Instituto Politécnico de Setúbal / Universidade Nova de Lisboa, Portugal
Andrzej Niesler, Wroclaw University of Economics and Business, Poland

Michel Occello, Université Grenoble Alpes, France
Matthias Olzmann, noventum consulting GmbH, Germany
Angel Pasqual del Pobil, Jaume I University, Spain
Agostino Poggi, University of Parma, Italy
Christoph Reich, NEC Laboratories America Inc., USA / TU Darmstadt, Germany
Fátima Rodrigues, Instituto de Engenharia | Politécnico do Porto, Portugal
José Rouillard, University of Lille, France
Minoru Sasaki, Ibaraki University, Japan
Ingo Schwab, Center of Applied Research (CAR) | University of Applied Sciences Karlsruhe, Germany
Hans-Werner Sehring, NORDAKADEMIE - University of Applied Sciences, Germany
Vasco N. G. J. Soares, Polytechnic Institute of Castelo Branco | Instituto de Telecomunicações, Portugal
Donald Sofge, Naval Research Laboratory (NRL), USA
Claudius Stern, FOM University of Applied Sciences for Economics and Management, Kassel, Germany
Sotirios Terzis, University of Strathclyde, UK
Mihaela Vranić, University of Zagreb, Croatia
Alexander Wijesinha, Towson University, USA
Yingcai Xiao, University of Akron, USA

International Journal on Advances in Intelligent Systems
Volume 18, Numbers 3&4, 2025

CONTENTS

pages: 79 - 89
Visualizing Similarities of Music from Different Instruments - A Novel Proposal to Project High-dimensional Music Features on a 2-dimensional Plane
Goutam Chakraborty, Iwate Prefectural University, Japan
Cedric Bornand, University of Applied Sciences HES-SO, Switzerland
Lokesh Ayyaswamy, Madanapalle Institute of Technology & Science, India
Lakshman Patti, Madanapalle Institute of Technology & Science, India
Praveen Kumar Reddy Sangati, Madanapalle Institute of Technology & Science, India
Subhash Molaka, Madanapalle Institute of Technology & Science, India

pages: 90 - 98
Dynamics of Communication Infrastructure in Transactive Energy Systems: A Systems Dynamics Model of LoRa Integration
Atefeh Zarei, Stevens Institute of Technology, United States
Evans Honu, Stevens Institute of Technology, United States
Mo Mansouri, Stevens Institute of Technology, United States
Philip Odonkor, Stevens Institute of Technology, United States

pages: 99 - 109
Emulating the Style of Johannes Brahms: Neural Network-Based Generation of Romantic Piano Music
James Doherty, Technological University Dublin, Ireland
Brendan Tierney, Technological University Dublin, Ireland

pages: 110 - 125
Extending the Composable Architecture Framework by Means of Normalized Systems Theory
Geert Haerens, Antwerp Management School, Belgium
Herwig Mannaert, University of Antwerp, Belgium

pages: 126 - 136
Enabling Kubernetes Workload Execution on Rootless HPC Systems with KSI: A Slurm Integration Framework
Jonathan Decker, Georg-August-University Göttingen, Germany
Mojtaba Akbari, GWDG, Germany
Ali Doosthosseini, Georg-August-University Göttingen, Germany
Sören Metje, Georg-August-University Göttingen, Germany
Aasish Kumar Sharma, Georg-August-University Göttingen, Germany
Julian Kunkel, Georg-August-University Göttingen, Germany

pages: 137 - 148
Designing Inclusive Immersive Experiences: Intergenerational Codesign of Heritage and Underwater Extended Reality for Older Adults
Marius N. Varga, University of Plymouth, United Kingdom
Oksana Hagen, University of Plymouth, United Kingdom
Rory Baxter, University of Plymouth, United Kingdom
Alejandro Veliz Reyes, University of Plymouth, United Kingdom
Ray B Jones, University of Plymouth, United Kingdom
Amir Aly, University of Plymouth, United Kingdom

Dena Bazazian, University of Plymouth, United Kingdom
Swen Gaudl, University of Gothenburg, Sweden

pages: 149 - 161
Direct Adaptive Input Matrix Estimation Approaches for Linear Time-Invariant Dynamic Systems with Known Inputs
Kevin Fuentes, Texas A&M University, United States
Mark Balas, Texas A&M University, United States
James Hubbard, Texas A&M University, United States

pages: 162 - 172
Why the Symbol Grounding Problem Matters in Virtual Reality: A Meme-Focused Solution Based on the Model Human Processor with Real-Time Constraints
Muneo Kitajima, Nagaoka University of Technology, Japan
Makoto Toyota, T-Method, Japan
Katsuko T. Nakahira, Nagaoka University of Technology, Japan

pages: 173 - 186
Transforming Conscious Goals into Unconscious Actions in Real-world Interactions: Real-world Use of Behavioral Ecological Memes via GOMS
Muneo Kitajima, Nagaoka University of Technology, Japan
Makoto Toyota, T-Method, Japan
Jérôme Dinet, Université de Lorraine, France
Katsuko T. Nakahira, Nagaoka University of Technology, Japan

pages: 187 - 198
Productive Simplification: Complexity Production by Multiple Reductive Simplification Procedures for Interpreting Multi-Layered Neural Networks
Ryotaro Kamimura, Tokai Univeristy, Japan

pages: 199 - 209
Leveraging Graph-Centric Case-Based Reasoning for Enhancing Monotonicity & AI Coherency
Steve Chan, VTIRL, VT/DE-CAIR, United States

pages: 210 - 237
Interstitial b-SHAP-Owen Amalgam for the Enhancement of Artificial Intelligence System-Centric Sequential Decision-Making
Steve Chan, VTIRL, VT/DE-CAIR, United States

pages: 238 - 259
Sound and Safety in the Age of Electric Vehicles: Pedestrian Insights from the EVA Project
Flaithrí Neff, Technological University of the Shannon, Ireland
Dónal Fitzpatrick, The National Disability Authority, Ireland
Gerald Craddock, The National Disability Authority, Ireland
Carly Cheevers, The National Disability Authority, Ireland
Aislinn Conway, The National Disability Authority, Ireland
Linda Butler, Hibernia College, Ireland

pages: 260 - 272
Leveraging Content and Structural Dynamics of Discourse for Rumor Verification

Gibson Nkhata, University of Arkansas, USA
Susan Gauch, University of Arkansas, USA

pages: 273 - 287
People First, Platform Second: Designing Hybrid Learning for Emotional and Educational Resilience in Unsafe Regions
Maria Eisenmann, Julius-Maximilians-Universität Würzburg, Germany
Nataliia Lazebna, Julius-Maximilians-Universität Würzburg, Germany
Kateryna Lut, National University "Zaporizhzhia Polytechnic", Ukraine

pages: 288 - 299
Consequences of EU Data Regulation for Smart Home Data Economy: Extending Analysis with a Case Study
Paul Seidel, Hochschule Mittweida University of Applied Sciences, Germany
Felix Fischer, Hochschule Mittweida University of Applied Sciences, Germany
Dirk Labudde, Hochschule Mittweida University of Applied Sciences, Germany

pages: 300 - 310
Defect Prevention Matrix Analysis by Generative AI
Shuichiro Yamamoto, IPUT in Nagoya, Japan

pages: 311 - 322
From Bias to Balance: Fairness-Aware Paper Recommendation for Equitable Peer Review
Uttamasha Anjally Oyshi, University of Arkansas, United States
Susan Gauch, University of Arkansas, United States

Visualizing Similarities of Music from Different Instruments - A Novel Proposal to Project High-dimensional Music Features on a 2-dimensional Plane

Goutam Chakraborty *
Iwate Prefectural University, Information Science, IPU, Iwate, Japan
Email: goutam@iwate-pu.ac.jp

Cedric Bornand†
University of Applied Sciences HES-SO
Yverdon-les-Bains, Switzerland
Email: cedric.bornand@heig-vd.ch

Lokesh Ayyaswamy‡, Lakshman Patti§, Praveen Kumar Reddy Sangati¶, Subhash Molaka‖
Department of Computer Science Engineering and Artificial Intelligence‡§¶‖
Madanapalle Institute of Technology & Science, Madanapalle, India‡§¶‖
Email: lokeshreddy2680@gmail.com‡, lakshmanpatti99@gmail.com§,
prawinreddy1909@gmail.com¶, molakasubhash@gmail.com‖

***Abstract*—We perceive music from various perspectives: the melody, the rhythm, the emotions or passions they evoke, the richness of sound, and how it correlates with the time of day (like Morning Raga) or with seasons (like Vivaldi's Four Seasons). This is a multimodal classification challenge for which correct data annotation is a difficult issue. In this work, we propose a method for visualizing audio signals from various musical instruments to find their variances, quantify their similarities, and distances. To facilitate visualizing, we project the audio data on a 2-dimensional plane such that the distances (dis-similarities) of the audio signals are preserved as much as possible. The appropriate tools (algorithms) for this task were identified by experimental analysis. The work is conducted in two stages: the first is audio feature extraction and compression, and the second is the projection of high-dimensional audio features on a two-dimensional plane using various unsupervised visualization techniques. The aim is to determine which feature compression and visualization tools can produce clearly separated clusters of audio signals. The features of the Short Time Fourier Transform (STFT) spectrograms extracted using Convolutional Neural Networks (CNN) provided the best compressed representations, facilitating clear separation and meaningful projection of the audio signals using visualization tools t-Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). The scatterplots of the samples achieved silhouette scores of 84% and 81%, respectively, ensuring clear groups for sounds generated from different instruments. We also experimented with UNet to find numerical vector representations of the spectrogram images. UNet could achieve silhouette scores of around 75%.**

***Keywords*-MFCC; STFT; Spectrogram; CNN; U-Net; t-SNE; UMAP.**

I. Introduction

This study analyzes audio signals from ten musical instruments, a combination of traditional Indian and Western ones. The features extracted from these signals are high-dimensional and projected onto a 2-dimensional plane to visualize similarities. The Indian instruments include flute, nadaswaram, and shehnai (wind-type); santoor and veena (string); and thavil and mridangam (percussion), while the Western ones are piano, guitar, and violin. Each instrument has distinct sound characteristics: wind instruments produce tones through vibration of air, string instruments through plucking or bowing, percussion by striking a tense diaphragm with a hand or a stick, and the piano by hammering strings via key presses.

This work builds upon our conference publication [1] expanding the number of instruments from six to ten and improving the visualization methodology.

The Fourier Transform (FT) [2] and the Fast FT (FFT) [2] were used to examine the audio signals. However, FFT cannot capture sequential information from the signal. Advances in speech processing introduced techniques such as STFT [3], Wavelet Transform (WT) [4], and Mel-frequency cepstral coefficients (MFCC) [5]. MFCC exploits the log scale of human audio perception and is widely used for audio signal analysis like speaker identification. As it uses small windows, the number of features increases linearly as the length of the audio signal increases. Using audio feature extraction methods and deep neural networks for compressing high-dimensional audio data enables us to achieve a relatively low-dimension representation of the audio signal, which is further used for visual representation as scatter plots on a 2-dimensional display. The overall plan is shown in Figure 1.

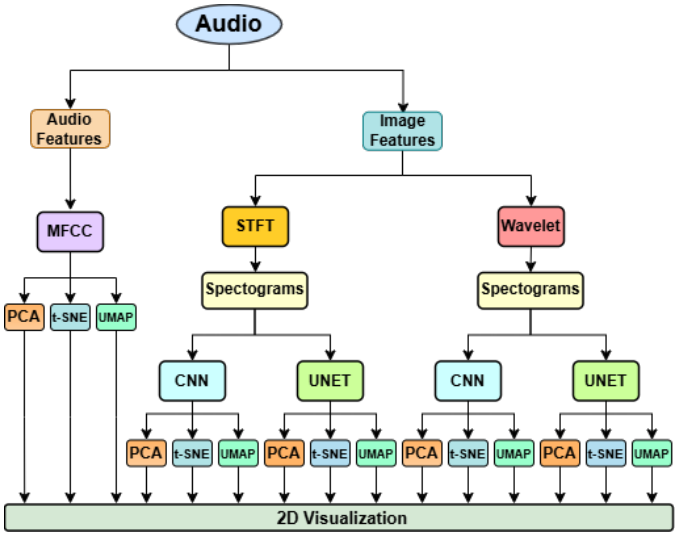


Figure 1. Overall plan for the Experiments.

The decorrelated MFCC features were extracted from the audio samples [5], using MFCC window lengths of 25 milliseconds. Even a few seconds of audio signal generate

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

a high-dimensional feature vector. To be able to capture the characteristic features of the audio samples, we used signals of length 30 seconds. We had to find ways to represent them as a relatively low-dimensional vector. Here, we used spectrograms and neural networks to extract features from the spectrogram images.

For spectral analysis, we used the STFT. We converted the STFT features into spectrogram images [3]. These spectrograms serve as visual representations of the musical features. We also converted the audio to wavelet coefficients using Morlet wavelets, and that is being transformed to a spectrogram image representation.

To extract features from spectrograms, we trained a CNN model [6], and a U-Net model [7] on STFT and wavelet spectrograms. As our sample classes are known, the deep neural network was trained as a supervised classifier. Features were taken from the output of the convolutional layers, which are inputs to the dense classification layer. In the U-Net model, features were extracted from the bottleneck layer at the end of the encoder.

The effectiveness of the proposed methods was validated through several experiments by projecting the features onto a two-dimensional plane [8]. Section III details how CNN and U-Net architectures extract features from STFT and WT spectrogram images. To visualize music signals on a two-dimensional plane, we used Principal Component Analysis (PCA) (a linear method), t-SNE, and UMAP (nonlinear methods).

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the methodology, including data collection and preprocessing, feature extraction, and the proposed solution. Section IV presents the experimental results and their analysis. Finally, Section V concludes the paper and discusses future research directions.

II. Related work

The previous works on the visualization of audio sample characteristics are discussed below.

The authors used three different datasets in their work reported in 2024 [9]. Two datasets with 10 classes and an augmented version of one (using pitch shifting, time-stretching, and added random noise) were used. MFCC features were extracted, and CNN and Recurrent Neural Network (RNN)-(Long Short Term Memory) LSTM models were trained. CNN performed better on smaller datasets, while RNN-LSTM excelled on larger ones.

In the work reported in 2020 [8], the authors experimented with audio data of 10 classes, extracting MFCC features. They visualized these high-dimensional features using PCA, t-SNE, Iso-Map, and SOM. t-SNE produced well-separated clusters. SOM showed slight separation, while Iso-Map failed to capture any meaningful clustering. The conclusion was that Iso-Map failed to work with this high-dimensional data.

In another work on the audio classifier, reported in 2020 [6], the authors used a public dataset and converted the audio signals into Mel power spectrograms. They applied two approaches to capture features: a CNN model trained from scratch and a pre-trained VGG19 model using transfer learning. Both models performed well. The CNN model trained from scratch slightly outperformed the VGG19 model.

III. Proposed Methods

This Section presents the workflow of our experiments, covering the collection of audio samples of ten musical instruments, preprocessing, feature extraction, dimensionality reduction, and finally 2-dimension projection.

A. Data Collection and Pre-processing

Audio samples were sourced from public platforms like YouTube and recorded media, ensuring that each sample captures the unique tonal and spectral qualities of the instrument without background noise or interference from other instruments.

We gathered 300 audio samples, 30 samples per instrument, using YTMP3 and converted them to MP3. We processed them with Clideo. Clips were segmented into 30–45 seconds, then converted to WAV, ensuring standard audio quality for audio signal analysis.

B. Feature Extraction

1) MFCC Feature Extraction: MFCC features are widely used in audio analysis for music, speech recognition, and speaker identification. Pre-processing involves standardizing samples to 30 seconds through padding or trimming, then sampling at 44,100 Hz to preserve the high-quality of audio samples.

The MFCC extraction process starts with splitting the 30-second audio into 25 ms non-overlapping frames (1,201 frames, each with 1,103 samples). The segmented audio will get distortions at the end points due to abrupt change. To smooth out the change, we apply Hanning window [10] to maintain smooth transition betweem frames. The Discrete Fourier Transform (DFT) transforms the signal to the frequency domain, capturing spectral characteristics. A Mel filter bank emulates human hearing by dividing the spectrum into 26 bands, reducing dimension while maintaining required information. A logarithmic transformation follows to constrict the dynamic range. The last step is Discrete Cosine Transform (DCT) to decorrelates Mel-spectral coefficients, retaining the first 13 MFCCs.

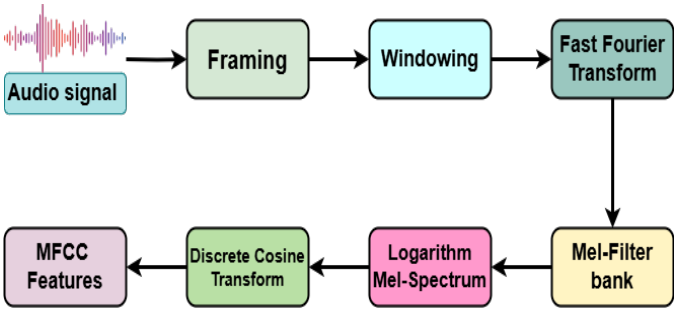


Figure 2. The process of MFCC Feature Extraction.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

The MFCC extraction process is shown in Figure 2. Each 30-seconds sample is converted into 13 MFCCs × 1,201 frames and flattened into a 1-D vector of 15,613 elements. MFCCs capture important audio characteristics, preserving tonal, timbral, and rhythmic features for analysis of the audio signal.

Finally we have 300 samples, 30 samples each from ten musical instruments. Each sample is represented by a vector of 15,613 elements, 13 MFCC values from 1201 frames.

2) STFT Spectrogram Generation: The audio sample of 30 seconds is segmented into windows of size 25 milliseconds. Each segment of 25 millisecond contains 1,103 sample points, because the sampling rate is 44,120. FFT is applied to extract the frequency content (frequency band and corresponding intensity) of the signal within the segmented window. These frequency domain representations are then concatenated to form the spectrogram images. The spectrogram is an image of the frequency spectrum where the intensity of a frequency band is represented as brightness [11]. Figure 3 shows the STFT spectrograms of different musical instruments, capturing the tonal and spectral characteristics of the instruments as images.

3) Wavelet Spectrogram Generation: The 30-second audio is segmented into 2-second chunks. We use a minimum frequency of 64 Hz and a maximum frequency as the Nyquist frequency which is half of the sampling rate 44100/2. We used 32 scales. The sampling rate high so as to be able to capture the delicate abrupt changes that happen with musical instruments. For voice a much lower sampling rate suffices. We apply Continuous wavelet transform using the Morlet wavelet. The Morlet wavelet is shown in Figure 4. Morlet wavelets are a product of a sine wave and a Gaussian function. The standard deviation of the Gaussian determines the window size. Different sine wave frequencies capture different frequency components of the audio signal. As mentioned, 32 scales were used. The wavelets convolve over the signal and gets the wavelet coefficients. Amplitudes are converted into decibels. Wavelet coefficients from all the segments are concatenated to represent the entire audio signal.

The wavelet spectrogram displays the frequency spectrum where the intensity of a frequency is converted into brightness. Each spectrogram represents the unique characteristic of an audio signal [12]. The wavelet spectrograms of the sample for each instrument are shown in Figure 5.

C. Projection of higher dimension into 2D

We thus have five sets of features - one obtained from MFCC, two sets from STFT spectrogram features and then using CNN and U-Net; and two sets of wavelet spectrogram features and then using CNN and U-Net. To visualize these features in 2 Dimension, we employed three different visualization techniques. Firstly, we visualized the extracted features using PCA by taking the first two principal components and projecting the samples on two dimension where the two principal eigen vectors are the basis [13]. Its capability for proper projection is limited due to its linearity restriction. We used two nonlinear tools for projection t-SNE [14] and UMAP [3]. t-SNE takes the distribution in the high dimension and maps it to two dimension using a non-linear method, by which data that are closer in high dimension become closer, and data that are far are pushed further. This helps to represent the data in tight clusters in 2-dimension. UMAP [3] on the other hand, works on manifold space. Distances between samples are measured in terms of probabilities, as a random walk on a Markov chain. Thus, the distances between two data points are measured on the manifold space, not in the Euclidean space. The two prominant embedding space directions were used as the two axes of 2-dimensional visualization space. UMAP could give excellent results, even when high dimension MFCC features were used.

D. Proposed Method

To map high-dimensional audio samples onto a two-dimensional plane, we employed three distinct visualization algorithms: PCA, t-SNE, and UMAP.

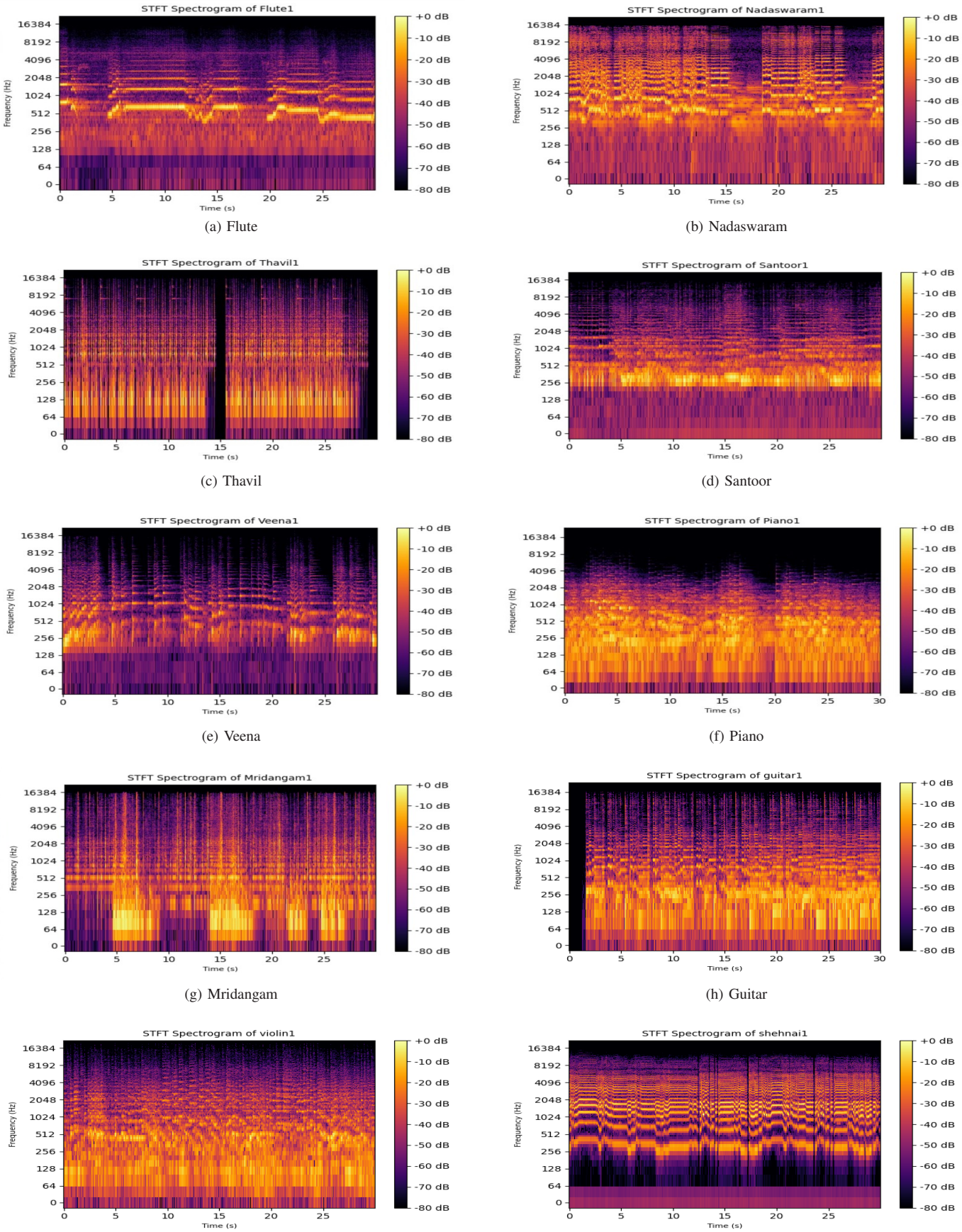
1) MFCC features onto 2D plane: MFCC features are extracted from audio signals, resulting in a high-dimensional data set with 15,613 dimensions for each 30-second music sample. In total, we have 300 samples from 10 different instruments. This dataset matrix of dimensions 300x15613 is the input to PCA, t-SNE, and UMAP to visualize the data on a 2-dimensional plane.

2) Feature Extraction using CNN: The STFT and WT spectrogram images of audio samples are used to train a Deep Neural Network (CNN) classifier model. The samples are labeled with the respective musical instruments. The output from the CNN layers are taken out as image features. The architecture of the CNN model consists of two convolutional layers, each succeeded by max - pooling layers, followed by a flatten layer and two dense layers for classification, one hidden and one output layer. The output from hidden layer is used as the compressed feature vector.

In Figure 6, the architecture of the CNN model utilized for training is shown. The model receives an STFT spectrogram images of size 400x600x3 as input. The first convolution layer comprises of 16 filters yielding an output of 400x600x16. Subsequently, a MaxPooling layer with a 2x2 kernel diminishes the size to 200x300x16. Following this, a second convolution layer with 32 filters is applied, and after executing 2x2 size max pooling, the output is further reduced to 100x150x32. The output is then flattened into a vector and processed through a dense layer classifier with a hidden layer of 64 nodes. The network is trained as a supervised classifier with 48,000 features as input, and 10 output nodes for 10 musical instruments. As mentioned, the hidden layer consists of 64 nodes. The extracted features are visualized using PCA, t-SNE, and UMAP algorithms.

For the wavelet spectrograms, which are of size 390×584×3, we use the same CNN architecture as for the STFT spectrograms. The only modification is that the first convolutional layer is configured with 20 filters. We then train a separate CNN model with this architecture and extract the features from the hidden dense layer for visualization.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) Flute (b) Nadaswaram

(c) Thavil (d) Santoor

(e) Veena (f) Piano

(g) Mridangam (h) Guitar

(i) Violin (j) Shenai

Figure 3. STFT spectrogram samples for ten musical instruments.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

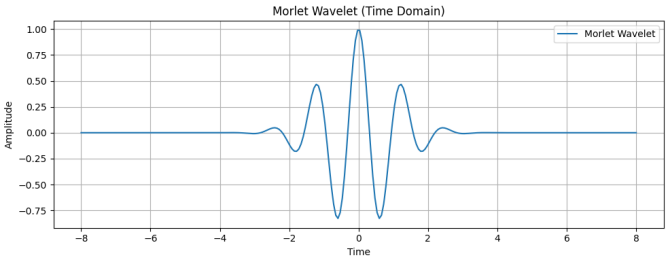


Figure 4. Morlet Wavelet In Wavelet Transform.

3) Feature Extraction using UNet: UNET, which was proposed for medical image segmentation, is used to compress the image features. The UNet architecture consists of encoder and decoder. The encoding part uses convolutional layers followed by max-pooling layers to extract features and reduce dimensions, while the decoding part employs upsampling layers to reconstruct the input. Essential compressed audio features are available in the bottom layer of the UNet.

In Figure 7, the architecture of UNET is shown. The STFT spectrograms of size 400×600×3 are input to the UNET model. Initially, two convolution layers with 32 filters are used, followed by max pooling with a pool size of 4×4. Next, two more Convolution layers with 64 filters are appended, followed by another MaxPooling operation. After that, two additional convolution layers are added, one with 128 filters and the next one with 64 filters, leading to the bottleneck layer, which captures the encoded representation of the input.

From the bottleneck layer, the features are upsampled using a transposed convolution operation with 64 filters of size 4×4. These up-sampled features are concatenated with corresponding layer from the encoding side. The process of convolution, upsampling, and concatenation is repeated for feature reconstruction. Finally, this process produces the reconstructed image. It is an unsupervised algorithm for encoding-decoding of images.

The elements from the bottleneck layer represent the input spectrogram image features. These values from the UNET bottom layer are the input to the three visualization software, namely PCA, t-SNE, and UMAP.

First, we train the UNet using data from all ten classes. The compressed features from the UNet bottleneck, from different classes, are superimposed in the feature space and fail to be presented as separate clusters on a 2-Dimensional plane. In the next experiments, we trained UNet separately with individual classes of samples. After this training, the features from the UNet bottleneck layer are used. The visualization algorithms projected them as isolated groups. Both STFT and wavelet spectrograms are used to train the UNET model and extract the features from the bottleneck layer, and the features are visualized.

IV. Experiments and Results

In this Section, we present the visualization results of MFCC features and the extracted features from STFT and WT spectrogram images using CNN and UNet.

A. Visualization of MFCC Features

To visualize high-dimensional MFCC features on a 2-dimensional plane, we used PCA, t-SNE, and UMAP. The resulting scatter plots of 300 music samples are shown in Figures 8, 9, and 10.

In Figure 8, the scatter plot shows that different instrument classes are mixed, with poor separation between samples from different musical instruments. All audio samples from different musical instruments are randomly placed on the 2-dimensional projection. This is because (1) the number of MFCC features is too large, and (2) PCA, being a linear method, cannot correctly project the data, which are on a curved manifold space in the MFCC feature space.

Both t-SNE and UMAP reveal distinct clustering patterns for different instruments based on their MFCC features. Piano consistently appears well-separated in both visualizations, reflecting its unique timbral profile. Thavil also shows clear separation in the t-SNE plot, whereas UMAP emphasizes tighter groupings among other instruments. Figure 10(UMAP) demonstrates a slightly better overall grouping of audio samples from different instruments than as Figure 9 (t-SNE). But neither of the two methods could achieve clear separation of all samples generated from 10 different instruments.

B. Visualization of Extracted features from CNN model

The STFT spectrogram and wavelet spectrogram are trained with CNN, and the features extracted from the hidden layer of the densenet classifier are now used as input to the visualization software. These features are much lower in number (64) than the MFCC features. These features extracted by CNN from both STFT and wavelet spectrograms are then input to PCA, t-SNE, and UMAP for visualization.

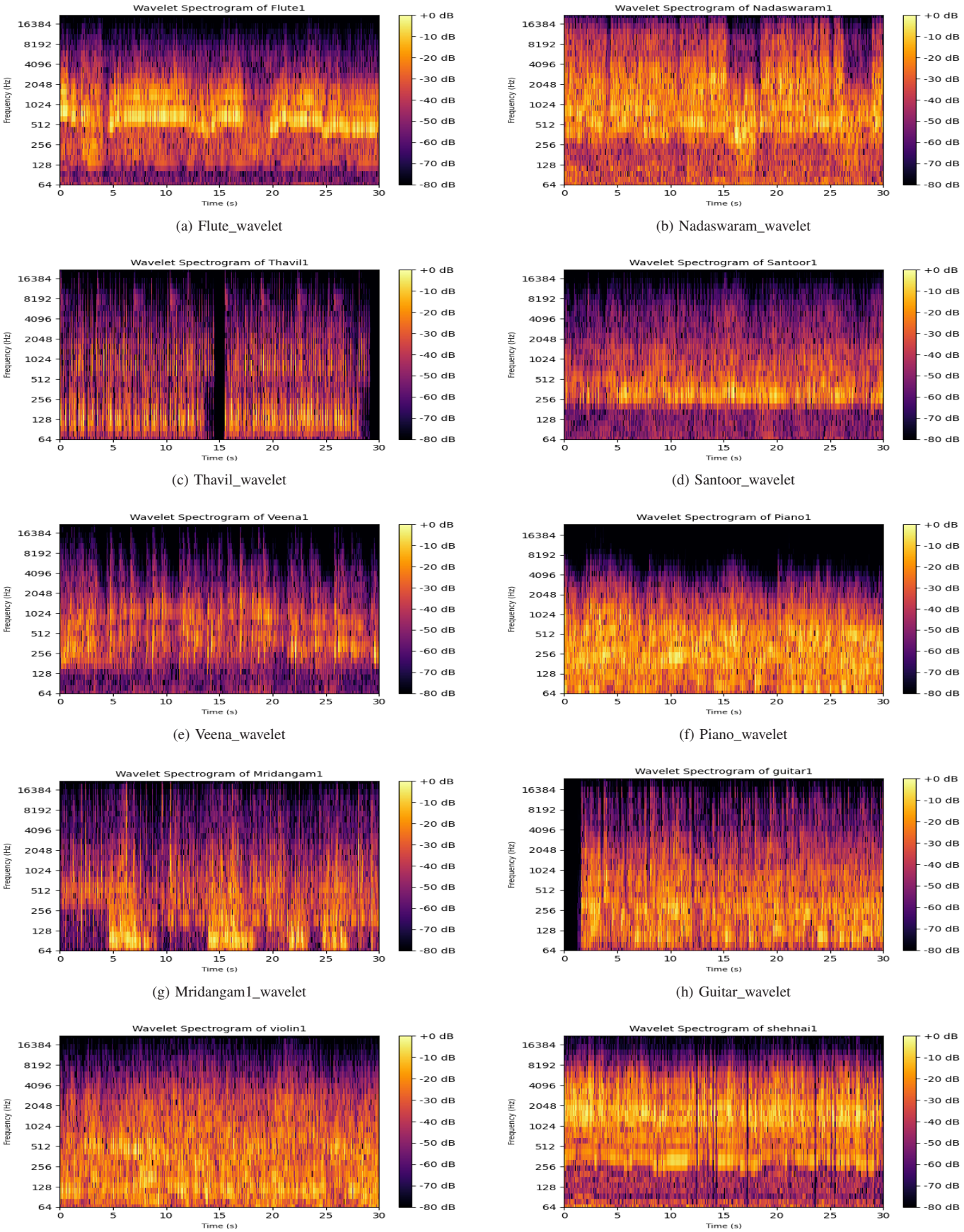
1) STFT spectrogram Features through CNN: The STFT spectrogram results are shown in Figures 11, 12 and 13.

In Figure 11, the PCA plot of CNN features displays moderately distinct groupings, but some overlaps persist. Clusters for classes like Santoor and Nadaswaram are visible, though not as tightly packed as in UMAP or t-SNE.

In Figure 12, the t-SNE representation of CNN features reveals distinct, compact clusters for all ten instruments. For instance, the shehnai and violin are far from the piano and mridhangam, suggesting meaningful class separation. By meaningful, we mean how we perceive the sounds from those 10 instruments - some perceived as more similar than others. This visualization confirms the CNN model's ability to learn feature representations that preserve inter-class differences effectively.

When UMAP was used for the 2-dimensional projection, as shown in Figure 13, highly compact clusters of samples from individual instruments were formed. The intra-cluster distances between violin, veena, thavil, guitar and mridhangam music

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) Flute_wavelet

(b) Nadaswaram_wavelet

(c) Thavil_wavelet

(d) Santoor_wavelet

(e) Veena_wavelet

(f) Piano_wavelet

(g) Mridangam1_wavelet

(h) Guitar_wavelet

(i) Violin_wavelet

(j) Shenai_wavelet

Figure 5. Wavelet spectrograms of different musical instruments.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

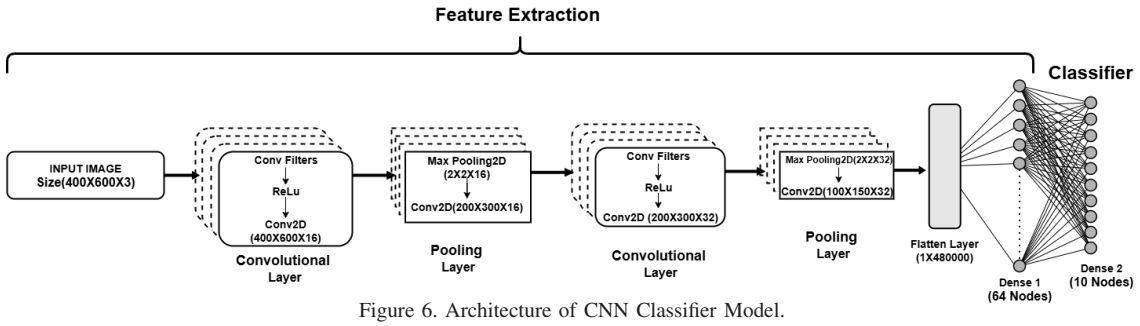


Figure 6. Architecture of CNN Classifier Model.

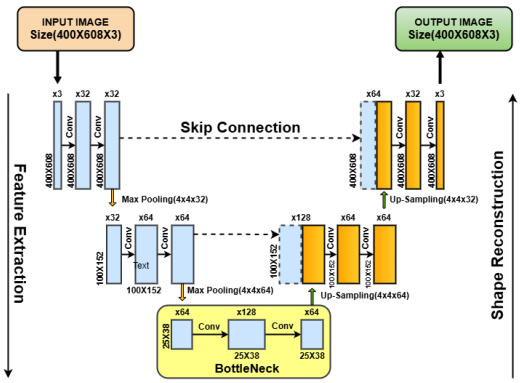


Figure 7. UNet Architecture.

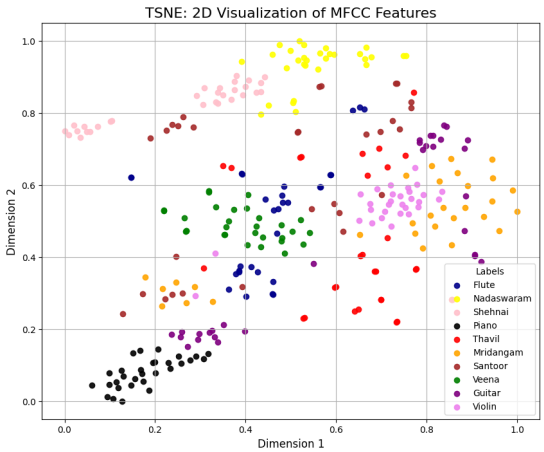


Figure 9. t-SNE visualization of MFCC features.

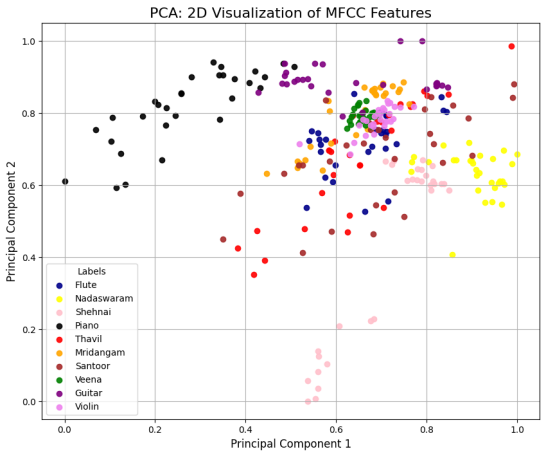


Figure 8. PCA visualization of MFCC features.

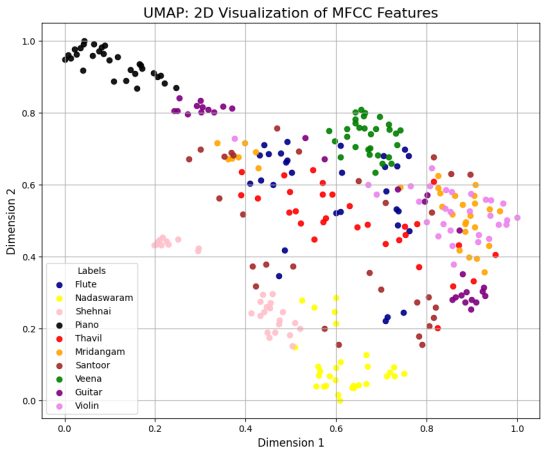


Figure 10. UMAP visualization of MFCC features.

samples are less. Compared to MFCC-UMAP, this visualization reflects a significant improvement in class separation, making UMAP effective and similar to projecting CNN features onto 2-dimensional space.

2) Wavelet spectrogram features using CNN: The results are shown in Figures 14, 15 and 16.

In Figure 14, PCA is used for 2-dimensional projection. As expected, the scatter plot of samples do not present clear clusters. Yet, we see samples from similar-sounding instruments are closely placed, and though there are no clear clusters, there is a vague separation visible, suggesting that the most significant variances in the features are somewhat aligned with class boundaries.

In Figure 15, t-SNE is used for visualization of samples from CNN features extracted from spectrograms. t-SNE is especially good in preserving the local structure of the data distribution in high dimension. The resulting visualization displays well-separated and compact clusters for each instrument class. Instruments such as the shehnai, thavil, and flute show clearly defined boundaries, indicating strong class discrimination.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

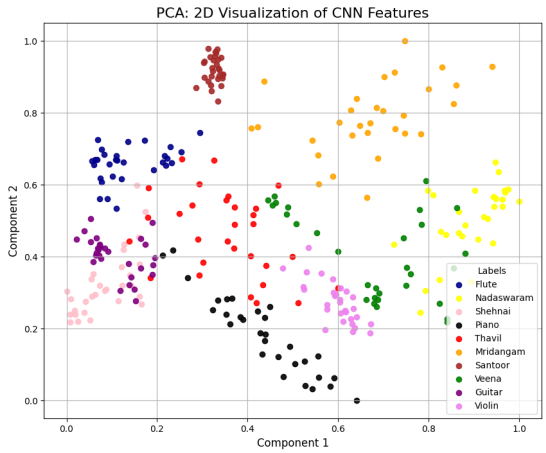


Figure 11. PCA visualization of STFT CNN features.

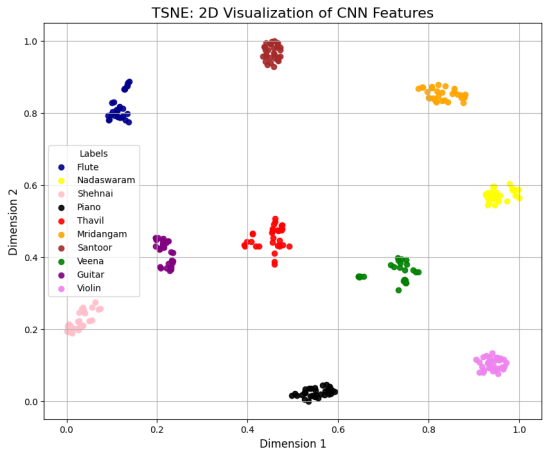


Figure 12. t-SNE visualization of STFT CNN features.

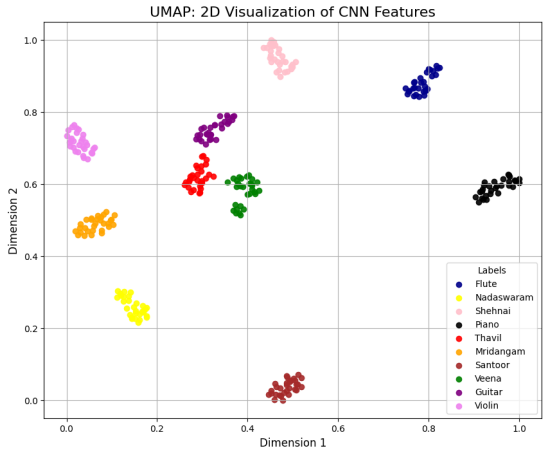


Figure 13. UMAP visualization of STFT CNN features.

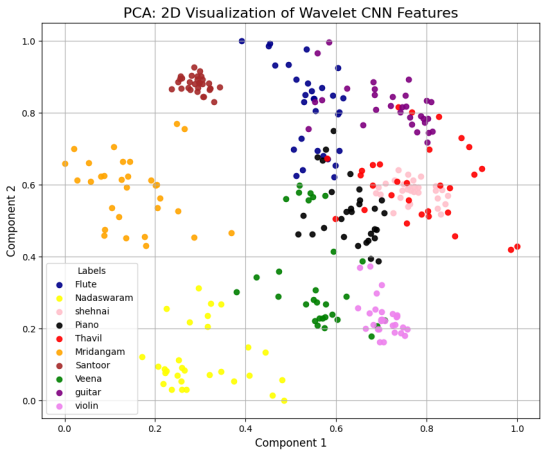


Figure 14. PCA visualization of wavelet CNN features.

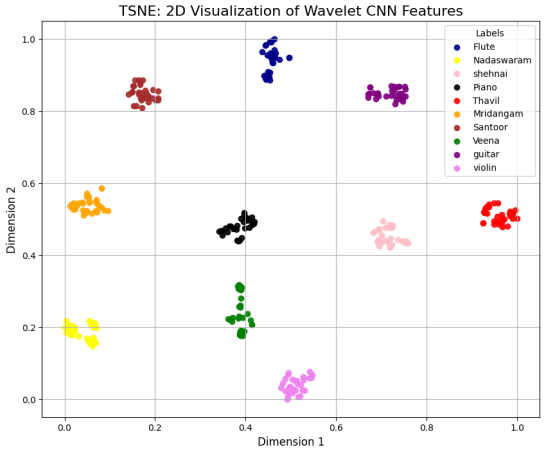


Figure 15. t-SNE visualization of wavelet CNN features.

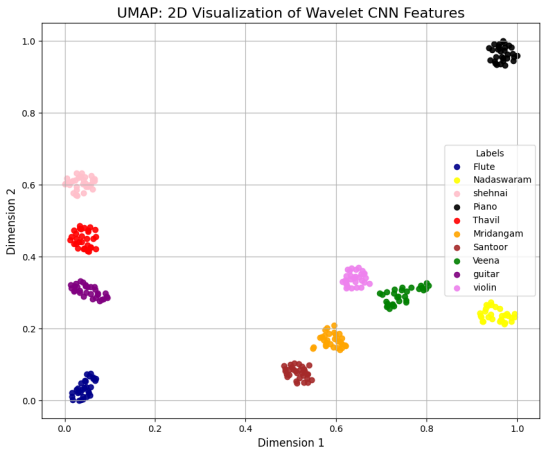


Figure 16. UMAP visualization of wavelet CNN features.

Unlike t-SNE, UMAP attempts to preserve both local and global structure. The UMAP projection, as shown in Figure 16, exhibits even better-separated clusters compared to t-SNE. Each instrument class occupies distinct regions in the scatter plot. This ensures that the Wavelet-CNN model has successfully captured both fine-grained and broad distinctions among instrument types.

C. Visualization Results of UNet Features

The STFT spectrograms and wavelet spectrograms were input into the UNet model, and features at the bottleneck layer were extracted. Thus, the original STFT features are compressed, and more abstraction is achieved at the UNet bottleneck. These compressed features are then used to visualize

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the data as a scatter plot on a 2-dimensional plane using PCA, t-SNE, and UMAP.

1) STFT spectrogram features through UNET: The STFT features extracted from the bottleneck are visualized, and results are shown in Figures 17, 18 and 19.

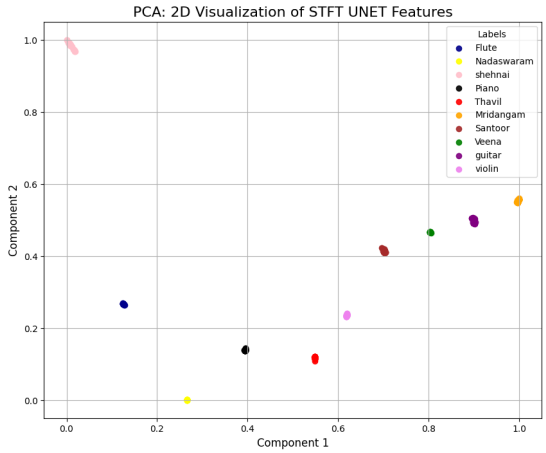


Figure 17. PCA visualization of UNET features.

In Figure 17, PCA projected samples forming compact clusters but widely separated. There are overlappings of samples from different instruments as well as splitting of samples from the same instrument. Flute forms a well-separated cluster. The other distinctive characteristics of Figure 17 is that samples appear to be compressed along the first principal component. As a linear dimensionality reduction method, PCA is less effective in separating the complex, non-linear relationships among UNET features.

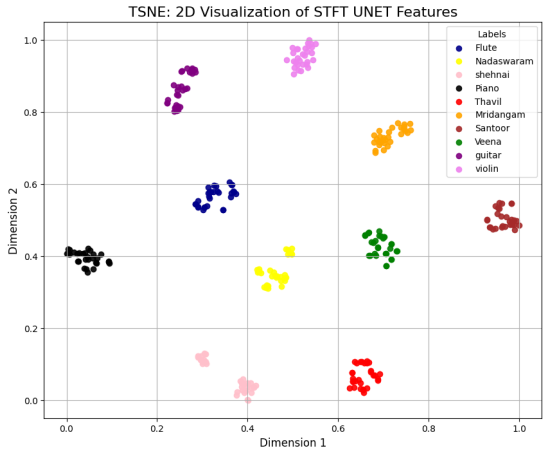


Figure 18. t-SNE visualization of UNET features.

In Figures. 18 and 19, the UMAP and t-SNE results on UNET features show clear, well-separated clusters for almost all instruments. UMAP reveals compact groupings, especially for flute, piano, nadaswaram, and violin, with low intra-cluster distances for thavil, piano, and shehnai. t-SNE also highlights distinct clusters. While minor dispersion exists in t-SNE, overall class separability is clear.

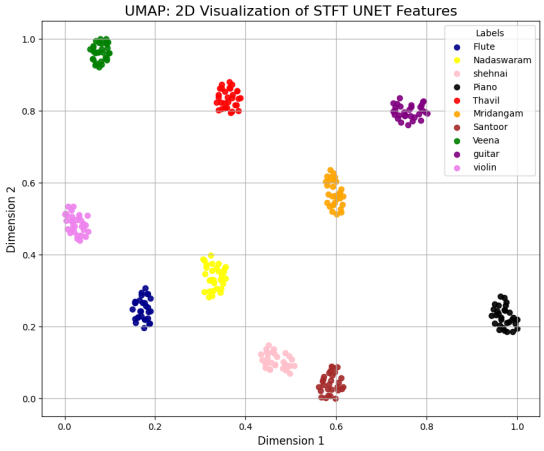


Figure 19. UMAP visualization of STFT UNet features.

2) Wavelet spectrogram features through UNET: The STFT features extracted from the bottleneck layer of UNET are used as input to the visualization tools. The results are shown in Figures 20, 21 and 22.

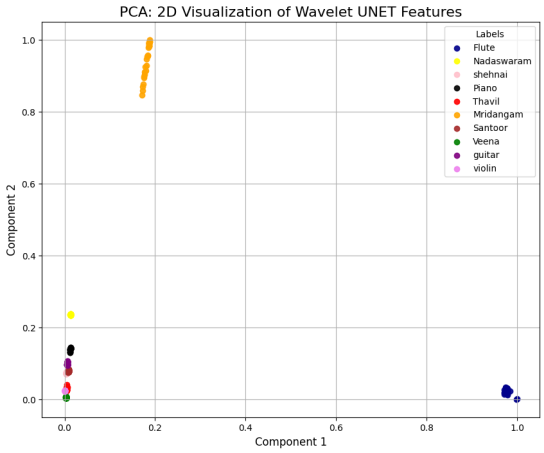


Figure 20. PCA visualization of Wavelet UNET features.

Figure 20 presents PCA results. Some instrument classes show partial separation, while a few samples from different classes overlapped. Despite limitations of the PCA being a linear projection algorithm, classes like Mridhangam and Piano form relatively distinct groups. But the shehnai and guitar music samples are overlapping.

In Figure 21, results using t-SNE are presented. We can see tightly clustered and well-separated groups. The strong local separation indicates that the model effectively distinguishes between instrument classes. This assures the improved ability of UNET to capture reliable feature representation.

Figure 22 shows the result when UMAP is used as the visualization tool. The figure shows distinct and compact clusters for each instrument class. The clear separation suggests that the combination of the Wavelet-UNET model effectively captures both local and global patterns in the data. Instruments like the flute, nadaswaram, and violin exhibit strong class distinction, which is what is perceived by human.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

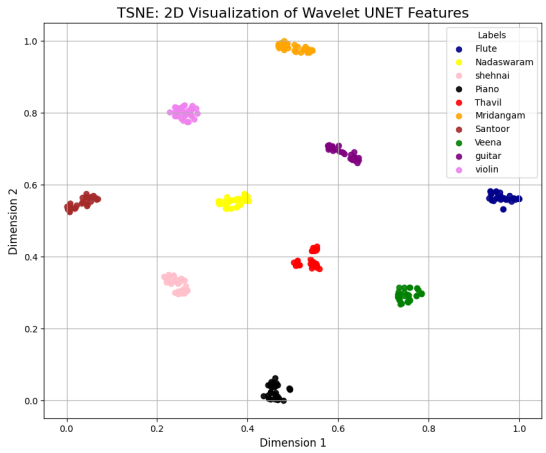


Figure 21. t-SNE visualization of Wavelet UNET features.

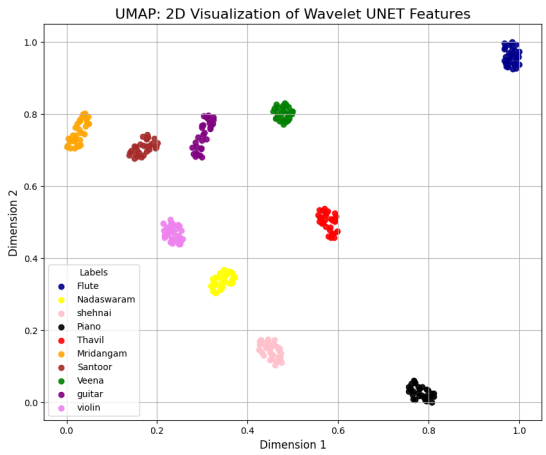


Figure 22. UMAP visualization of Wavelet UNet features.

TABLE I. COMPARISON OF VISUALIZATION TECHNIQUES BASED ON SILHOUETTE SCORES

Technique \ Features	MFCC	STFT (CNN)[a]	STFT (UNet)[b]	Wavelet (CNN)[a]	Wavelet (UNet)[b]
PCA	7.30	30.32	73.74	31.99	69.52
t-SNE	14.22	**84.72**	**78.09**	**84.35**	**78.83**
UMAP	**19.18**	78.66	75.05	79.20	74.47

The Silhouette scores, which are the ratio of interclass and intraclass distances, are displayed in Table I. High silhouette score means better separation between more compact clusters.

PCA demonstrates moderate performance for MFCC features and STFT spectrogram features using CNN but performs significantly better for the STFT features using UNET. t-SNE and UMAP outperform PCA in all experiments, with t-SNE achieving the highest silhouette scores. Both t-SNE and UMAP show similar performance for the STFT features using UNET, indicating their suitability for high-dimensional feature visualization. t-SNE gives a better compression closer samples are more tightly packed, and far-away samples are pushed farther away.

V. CONCLUSION AND FUTURE WORK

This study aims to find the correct tools to successfully visualize complex audio signals from musical instruments using machine learning and deep learning techniques. MFCC features, STFT, and wavelet spectrogram features were extracted to visualize the music samples in a two-dimensional plane. Three visualizations tools namely PCA, t-SNE, and UMAP are used. STFT features and wavelet coefficients were converted to spectrograms. Deep learning models and UNet were used to obtain a compressed version of the spectrogram image features. t-SNE and UMAP gave the best results, showing well-separated clusters. Though t-SNE gives better silhouette scores, UMAP projections are more akin to human perception. This is because t-SNE works locally whereas UMAP works on the manifold space which preservea the global information about sample distances.

MFCC features are proven to be the best for applications like speaker identification. But, for our application it is not. Instead of using MFCC features, concatenated spectrograms from sequential segments form an image for the whole musical sample. Analyzing those images are a better way to get distinct, separated clusters. The CNN model is sensitive to hyper parameters. The UNet model is robust and works better at all times. We train and extract the features, and it gives distinct separation between clusters. It is difficult to quantify the correctness of the results as far as human perception is concerned. For further investigation, we will

- Find the first few eigenvalues to check how fast the eigenvalues are diminishing and how that is reflected when the data is projected on the plane of the first two eigenvectors.
- Compare the interclass distances resulting from three different visualization algorithms and whether the relative distances from different methods are similar or not.
- Implement SOM as a tool for 2D visualization.

We will also extend this work for music generation, combining music generated by different instruments.

REFERENCES

[1] Goutam Chakraborty, Cedric Bornand, Lokesh Ayyaswamy, Subhash Molaka, Praveen Reddy, and Lakshman Patti, "Visualizing proximity of audio signals from different musical instruments: A two step approach", in *DBKDA 2025: The Seventeenth International Conference on Advances in Databases, Knowledge, and Data Applications*, Accessed: March 9, 2025, Porto, Portugal: IARIA, Mar. 2025, pp. 25–31, ISBN: 978-1-68558-244-9.

[2] Meinard Müller, "The Fourier Transform in a Nutshell", in *Fundamentals of Music Processing*. Aug. 2015, pp. 39–57, ISBN: 978-3-319-21944-8.

[3] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction", *arXiv preprint arXiv:1802.03426*, 2018.

[4] Anca Popescu, Inge Gavat, and Mihai Datcu, "Wavelet analysis for audio signals with music classification applications", in *2009 Proceedings of the 5-th Conference on Speech Technology and Human-Computer Dialogue*, 2009, pp. 1–6. DOI: 10.1109/SPED.2009.5156187

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[5] S. Memon M. A. Hossan and M. A. Gregory, "A novel approach for MFCC feature extraction", *2010 4th International Conference on Signal Processing and Communication Systems*, pp. 1–5, 2010. DOI: 10.1109/ICSPCS.2010.5709752

[6] Boyang Zhang, Jared Leitner, and Samuel Thornton, "Audio recognition using mel spectrograms and convolution neural networks", *Noiselab University of California: San Diego, CA, USA*, 2019.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation", in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[8] Tamás Pál and Dániel T Várkonyi, "Comparison of Dimensionality Reduction Techniques on Audio Signals", in *ITAT*, 2020, pp. 161–168.

[9] Karim Mohammed Rezaul, Md Jewel, Md Shabiul Islam, KNEA Siddiquee, N Barua, MA Rahman, et al., "Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models", *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, pp. 37–53, 2024.

[10] Hann function, *Hann function — Wikipedia, the free encyclopedia*, Online availability verified on 2 December, 2025, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Hann_function

[11] Eva Wesfreid, "STFT Time-Frequency Visualization: Application to Sound Signals", *Image Processing On Line*, 2013, Preprint, September 18, 2013. Online availability verified on 2 December 2025, ISSN: 2105-1232. [Online]. Available: https://dev.ipol.im/~eva/STFT.pdf

[12] Dongyu Wang, Canghong Shi, Xiaojie Li, Kai Peng, Xianhua Niu, and Sani M. Abdullahi, "Audio Manipulation Detection Based on Wavelet Spectrogram and Multidimensional Feature Fusion with Dual-Channel CNN", *SSRN Electronic Journal*, 2024. DOI: 10.2139/ssrn.5106730

[13] Samuele Battaglino and Erdem Koyuncu, "A generalization of principal component analysis", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 3607–3611.

[14] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE", *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Dynamics of Communication Infrastructure in Transactive Energy Systems: A Systems Dynamics Model of LoRa Integration

Atefeh Zarei, Evans Honu, Mo Mansouri, and Philip Odonkor*
Department of Systems and Enterprises
Stevens Institute of Technology
Hoboken, USA
email: azarei1@stevens.edu, ehonu@stevens.edu, mmansour@stevens.edu, podonkor@stevens.edu

***Abstract*—Distributed energy resources (DERs) introduce variability and bidirectional power flows that traditional centralized grids struggle to manage. Transactive Energy Systems (TES) offer a decentralized framework for balancing supply and demand through market-based coordination, but their deployment is constrained by communication infrastructure. This paper investigates Long Range (LoRa) technology as a TES communication solution using a system dynamics approach that models stakeholder interactions and feedback mechanisms. The analysis identifies sixteen feedback loops—eleven reinforcing and five balancing—that shape LoRa-TES integration. Critical technical relationships were empirically validated through packet-level network simulations. Results show that adaptive LoRa configurations reduce network-wide energy consumption by more than 99% compared to fixed-parameter baselines and sustain packet delivery ratios above 90% at low–moderate node counts, but also reveal a scalability ceiling of roughly 800–1000 nodes imposed by congestion. These findings highlight LoRa's potential as an energy-efficient communication layer for small to medium TES deployments, while underscoring the need for hybrid architectures and supportive policy frameworks to overcome scalability and interoperability barriers.**

***Keywords - Transactive Energy Systems; LoRa Technology; System Dynamics; Distributed Energy Resources; Energy Markets*.**

I. Introduction

This paper extends our preliminary work presented at ENERGY 2025 [1], which introduced a systems dynamics framework for analyzing LoRa integration in transactive energy systems. The rapid growth of DERs is transforming the operation of electric power systems. Solar panels, wind turbines, and battery storage create bidirectional power flows and introduce supply–demand variability that centralized grid infrastructure was never designed to manage [2]. These challenges are particularly pronounced in developing regions, where expanding electricity demand must be met alongside ambitious sustainability goals [3].

TES have emerged as a promising response. Instead of relying on centralized control, TES establish decentralized marketplaces where supply and demand are balanced in real time through price signals [4]–[6]. Early demonstrations suggest this approach can improve efficiency and reduce costs [7]. Yet the scalability of TES is constrained by a fundamental bottleneck: the communication infrastructure on which they depend.

Current communication technologies struggle to satisfy three essential TES requirements: energy-efficient operation across geographically dispersed resources, reliable long-range connectivity, and secure exchange of transaction data [8], [9]. These limitations hinder deployment, particularly in rural regions where distributed generation is most needed [10], [11].

Low-Power Wide-Area Network (LPWAN) technologies offer a potential path forward. Among them, LoRa has attracted significant attention for its ability to provide long-range connectivity, low power consumption, and strong signal penetration—characteristics that align well with TES operational needs [12]–[14]. Still, the complex ways in which LoRa's technical properties interact with TES performance remain insufficiently understood [15]. Without such understanding, system architects lack the evidence needed to design robust and scalable implementations.

This paper addresses that gap by analyzing LoRa-enabled TES through a system dynamics perspective. We develop a causal model that traces how communication technology choices propagate through reliability, security, and scalability outcomes [16]. The analysis identifies sixteen feedback loops—eleven reinforcing and five balancing—that structure system behavior, and it situates these dynamics within a stakeholder framework linking technology providers, producers, consumers, and regulators. Building on this model, we employ network simulation to validate critical technical mechanisms, demonstrating both the efficiency gains from adaptive communication and the scalability limits imposed by congestion. Finally, we draw from these insights to highlight leverage points and practical barriers that must be addressed for effective TES deployment.

The current journal version provides substantially deeper analysis across multiple dimensions compared to the preliminary conference paper. We expand the feedback loop analysis to examine sixteen interconnected mechanisms in detail, compared to the initial examination in the conference version. We introduce a comprehensive stakeholder interaction framework that maps the relationships among technology providers, energy producers and consumers, government agencies, and environmental organizations. The environmental implications receive expanded treatment, including detailed analysis of how LoRa enables weather forecasting for renewable energy optimization and reduces electronic waste through extended device lifespans. We also provide network simulation results

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

that validate key technical mechanisms and quantify performance trade-offs. Finally, this version includes extensive discussion of implementation challenges, particularly LoRa's data rate limitations, and offers practical recommendations for hybrid communication architectures that combine LoRa with complementary technologies.

The remainder of the paper is organized as follows. Section II reviews TES communication requirements and LPWAN capabilities. Section III presents the system dynamics methodology. Sections IV and V detail model development, results and validation. Section VI discusses stakeholder and design implications, and Section VII concludes with recommendations for future research.

II. Literature Review

The traditional power grid is undergoing a fundamental transformation, shifting away from its historical one-way electricity delivery from centralized plants to end consumers. This change is driven by environmental concerns, technological advancements, and evolving consumer needs [17], [18]. A major catalyst in this transition is the integration of DERs, such as rooftop solar panels, small wind turbines, and battery storage systems, which are positioned closer to consumption points and enable localized energy production [2]. However, integrating DERs into the grid requires a market framework that allows dynamic energy exchange, leading to the emergence of TES.

TES facilitate decentralized electricity trading through automated market mechanisms, empowering prosumers—entities that both consume and produce energy—to optimize their energy usage and trade excess generation at competitive prices [4], [5], [19], [20]. Unlike traditional grid structures, where excess energy must be sold to the main grid under regulated tariffs, TES allows prosumers to engage directly with local consumers, promoting flexibility and efficiency [21]. The effectiveness of TES hinges on several interconnected components: microgrids, which can operate independently or in conjunction with the main grid; smart meters, which provide real-time monitoring and automated trading capabilities; and energy management systems, which optimize energy flow and market transactions [8]. Despite these advantages, TES implementation presents significant challenges, particularly in communication infrastructure, data security, privacy, system scalability, and seamless integration with existing grid operations [10], [11], [15].

The communication backbone of TES plays a pivotal role in enabling secure and efficient transactions. Various technologies have been explored to facilitate information exchange among market participants, including WiFi and cellular networks. While WiFi provides a widely available and cost-effective solution, its limitations—such as range restrictions, high power consumption, and susceptibility to interference—hinder its scalability, particularly in dense urban environments [22]. Cellular networks, on the other hand, offer extensive coverage and reliable data transfer but come with high operational costs and significant energy demands, making them less ideal for large-scale TES deployments [23], [24]. These limitations necessitate alternative communication technologies that can balance energy efficiency, cost, and performance.

LoRa technology has emerged as a promising alternative for TES communication networks. Designed for Internet of Things applications, LoRa operates on a LPWAN protocol, offering long-range connectivity with minimal power consumption and strong interference resistance [12], [25]. Studies have demonstrated its advantages over traditional IoT communication technologies, highlighting superior energy efficiency, scalability, and security features [13], [14], [26]. However, trade-offs exist between reliability and energy efficiency, as efforts to enhance data transmission reliability often lead to increased energy consumption of end devices [27], [28]. Research on LoRa scalability has shown that network performance can be improved by adding gateways or adopting dynamic transmission strategies, yet increased deployment density may lead to interference and reduced coverage probability [14], [29], [30].

The integration of LoRa into TES introduces complex interactions among technical, economic, and social factors, necessitating a holistic analytical approach. Systems thinking provides a framework for understanding these interdependencies, emphasizing feedback mechanisms and causal relationships within the energy ecosystem [16]. Previous studies applying this approach to energy systems have demonstrated its value in identifying unintended consequences and optimizing the performance of emerging technologies [16]. By leveraging systems thinking methodologies, such as causal loop analysis, researchers can assess how LoRa influences key TES attributes—efficiency, security, scalability, reliability, and cost-effectiveness—while considering the broader implications of communication technology choices. This perspective is crucial for developing strategic implementation plans and mitigating potential challenges in real-world TES deployments.

III. Methodology

This section presents the methodological framework employed to analyze the integration of LoRa technology within Transactive Energy Systems. The methodology is structured around a systems dynamics approach that progressively builds from conceptual foundations to detailed causal models, enabling a comprehensive examination of the complex interactions between communication infrastructure and energy market dynamics.

A. A Systems Dynamics Approach to Transactive Energy System Analysis

The intricate structure of TES necessitates an analytical framework that captures both direct interactions and emergent system-wide behaviors. Traditional analytical approaches often fall short in accounting for the interdependencies among market participants, communication networks, regulatory frameworks, and technological infrastructures. Systems dynamics offers a robust alternative, enabling a holistic examination of TES by mapping causal relationships, feedback loops, and evolving system states over time [16]. This approach not only

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

facilitates a deeper understanding of TES operations but also provides a structured methodology for assessing the impact of integrating LoRa technology into these systems.

Our analysis unfolds across three progressive stages: defining system boundaries, identifying stakeholder interactions, and constructing causal loop models. Each stage builds upon the previous, culminating in a comprehensive evaluation of how communication infrastructure—specifically LoRa—shapes TES performance, scalability, and sustainability.

B. Defining System Boundaries

Establishing clear system boundaries is fundamental to understanding the scope and limitations of the analysis. In this study, the TES ecosystem is conceptualized as a network of interconnected entities, including energy producers, consumers, prosumers, grid operators, and technology providers. These actors operate within a landscape shaped by both technical components—such as DERs, microgrids, and communication networks—and non-technical factors, including regulatory policies, economic incentives, and market dynamics, as illustrated in Figure 1.

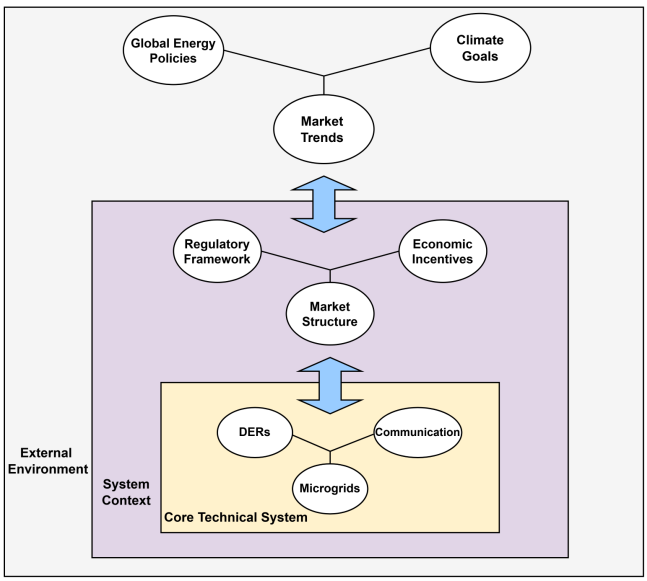


Figure 1. TES boundaries

While TES function within broader global energy trends, our boundary definition focuses on localized system interactions, excluding external influences such as international energy policies or technological developments beyond the direct operational scope of TES. This refinement ensures a targeted investigation into the role of communication technologies while maintaining analytical depth.

C. Stakeholder Interactions and System Dynamics

The complexity of TES arises from the intricate network of stakeholder relationships that shape system behavior and performance. These interactions, depicted in Figure 2, extend beyond direct transactional exchanges, encompassing regulatory influence, technological dependencies, and sustainability considerations. Understanding these dynamics is essential for identifying intervention points, anticipating systemic responses to change, and fostering resilience in energy markets. The

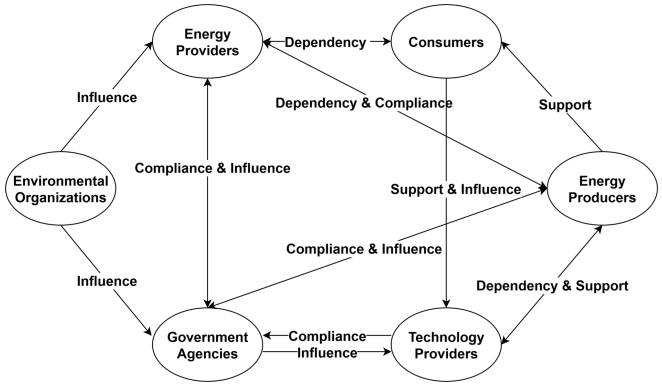


Figure 2. Stakeholder interest map in TES with LoRa integration

diagram illustrates the complex web of relationships between key stakeholders in TES. Arrows indicate primary interaction types: dependency (resource or service reliance), support (financial or operational assistance), compliance (regulatory or standard adherence), and influence (impact on decision-making or behavior).

At the foundation of TES, the interdependence between Energy Providers and Consumers forms a bidirectional relationship that is fundamental to system stability. Energy Providers ensure a reliable supply of electricity, while Consumers generate the demand that sustains economic viability. This relationship is further shaped by Energy Producers, who serve as both suppliers and market participants. Their operations are driven by demand fluctuations, policy frameworks, and technological advancements that influence production capacity and energy distribution.

Environmental Organizations exert a significant shaping force on the TES landscape, primarily through their advocacy for sustainable energy practices and stringent environmental regulations. Their influence is particularly pronounced in their interactions with Government Agencies and Energy Providers, where they drive policy decisions related to renewable energy integration, emissions reductions, and sustainability reporting. These pressures translate into regulatory mandates that compel stakeholders to align operational strategies with broader environmental objectives.

Government Agencies function as both regulators and facilitators within TES, enforcing compliance while simultaneously shaping market conditions through policy interventions. Their engagement spans multiple stakeholders, including Energy Providers, Producers, and Technology Providers, ensuring that system-wide objectives such as reliability, equity, and sustainability are maintained. This dual role enables Government Agencies to mediate competing interests, balancing economic

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

viability with regulatory imperatives.

Technology Providers play a crucial role in sustaining and advancing TES, offering infrastructure solutions that facilitate energy transactions, enhance grid reliability, and ensure regulatory compliance. Their relationship with Government Agencies is particularly dynamic, as evolving policy frameworks necessitate continuous technological adaptation. Moreover, their support relationships with Energy Producers underscore the increasing reliance on digitalization and smart grid solutions in energy management.

These interconnected relationships give rise to critical feedback loops that reinforce or modify system dynamics. Notable among these are:

- The Compliance-Driven Innovation Loop: Energy Providers, under regulatory pressure from Government Agencies, seek advanced solutions from Technology Providers, leading to innovations that enhance reliability and compliance.
- The Market Development and Adoption Loop: Consumer preferences influence Energy Producers, who in turn engage with Technology Providers to adopt solutions that meet emerging market demands, shaping the trajectory of technological advancements.
- The Sustainability Influence Loop: Environmental Organizations exert pressure on both Government Agencies and Energy Providers, driving legislative changes that enforce sustainability measures, thereby altering operational and investment priorities.

Recognizing and analyzing these feedback mechanisms is essential for stakeholders aiming to navigate TES complexities effectively. Identifying points of leverage within these interactions can facilitate targeted interventions that enhance system efficiency, promote technological innovation, and support the integration of emerging solutions such as LoRa technology. Furthermore, a nuanced understanding of stakeholder dynamics enables proactive management of potential resistance to change, ensuring that transitions toward sustainable and resilient energy systems occur smoothly and equitably.

D. Causal Loop Modeling and System Behavior

Building upon stakeholder interactions, we construct causal loop diagrams to map the interdependencies among key system variables, offering a structured perspective on system behavior. This iterative modeling process reveals two fundamental feedback mechanisms: reinforcing loops, which amplify trends within the system, and balancing loops, which introduce constraints that stabilize outcomes. By delineating these relationships, we gain insight into how the integration of LoRa technology influences TES across multiple dimensions.

One crucial aspect is cost dynamics, where energy consumption patterns and infrastructure requirements shape economic feasibility. System performance, particularly communication reliability and security, emerges as another critical factor, influencing the overall resilience of TES. Environmental considerations also come into play, as energy efficiency and resource utilization determine sustainability outcomes. Meanwhile, implementation challenges—including technical requirements and integration complexities—affect the practicality of deploying LoRa technology within existing frameworks.

The resulting causal loop diagram provides a visual representation of these interactions, highlighting leverage points where targeted interventions can optimize TES performance. By capturing the interconnected nature of TES components, this model facilitates scenario analysis, allowing stakeholders to anticipate system behavior under different implementation strategies.

Adopting a systems dynamics perspective in TES analysis offers several key advantages. It uncovers hidden dependencies that conventional assessments might overlook, bringing potential unintended consequences to light before large-scale deployment. Additionally, it provides a structured framework for evaluating trade-offs between technological capabilities, economic feasibility, and regulatory constraints. More importantly, this approach enables an integrated examination of how communication technology choices shape the long-term evolution of TES, equipping decision-makers with the insights necessary to develop resilient, efficient, and sustainable transactive energy markets.

IV. Results

Our systems dynamics analysis, illustrated in Figure 3, reveals a complex web of interactions between LoRa technology integration and TES performance. Through careful examination of the causal loop diagram, we identify eleven reinforcing mechanisms (R1-R11) and five balancing loops (B1-B5) that potentially shape system behavior across multiple dimensions. These interconnected feedback mechanisms offer deep insights into how communication technology choices influence system evolution. Our theoretical model identifies several key feedback mechanisms. The reinforcing loops include energy consumption reduction through low-power operation (R1), adaptive power control optimization (R3), infrastructure cost savings through long-range capabilities (R4), enhanced DER monitoring effectiveness (R6), e-waste reduction through extended battery life (R7), system-wide communication improvements (R8), weather forecasting integration for renewable optimization (R9), system reliability through interference-resistant communication (R10), and additional efficiency and operational loops (R2, R5, R11). The balancing loops encompass security vs adoption trade-offs (B1), authentication overhead effects (B2), network congestion limitations (B3), operational burden scaling effects (B4), and implementation cost constraints (B5).

The integration of LoRa technology influences system efficiency through several interconnected pathways. At the core, continuous monitoring of distributed energy resources creates a primary reinforcing loop (R6) that enhances system performance. When energy consumption decreases through LoRa's low-power operation, devices maintain longer monitoring periods without battery replacement. This benefit is amplified by early fault detection capabilities, creating a positive feedback loop that partially offsets the maintenance burden

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

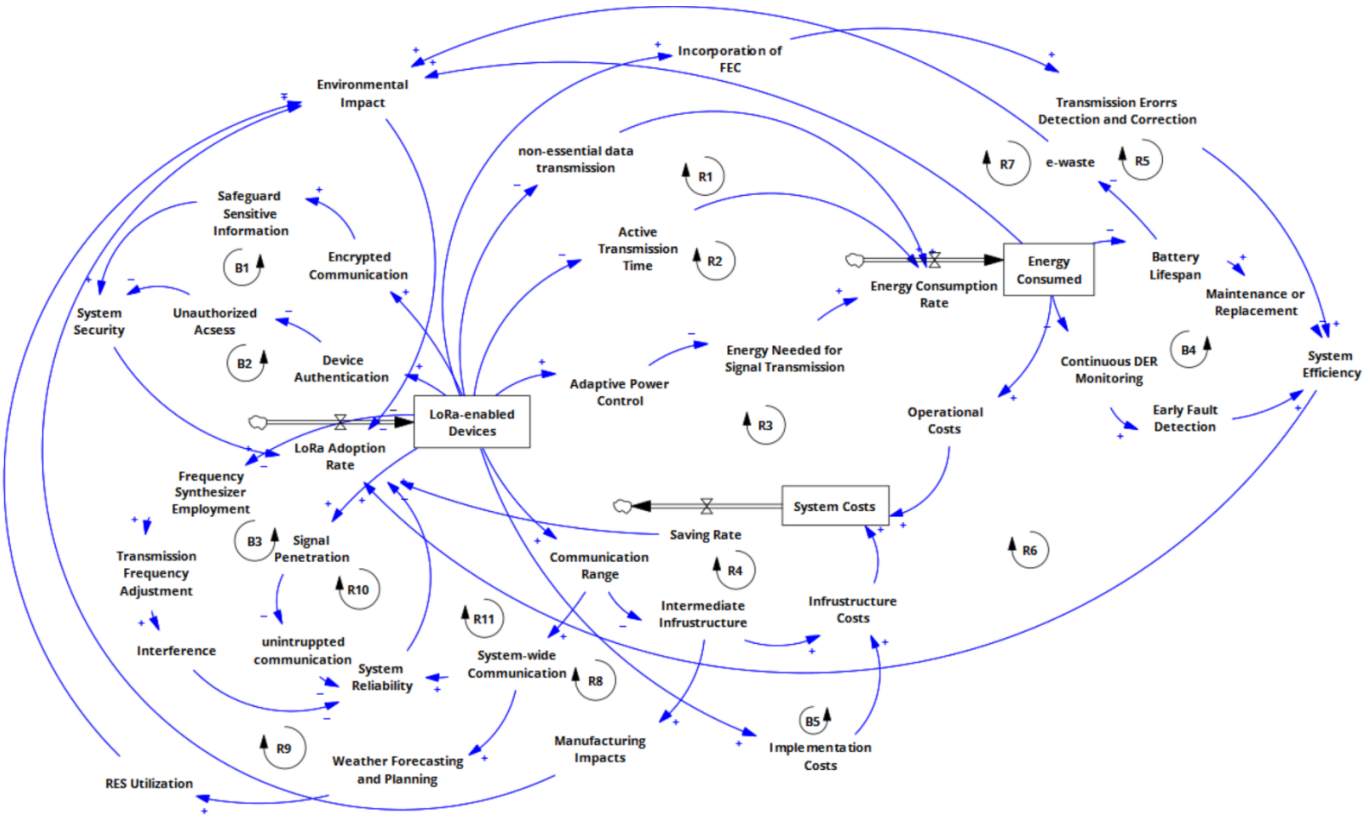


Figure 3. Causal loop diagram of impact of using LoRa on energy consumption behavior

typically associated with large-scale deployments. However, as system scale increases, maintenance requirements (B4) create a counterbalancing effect, suggesting the need for predictive maintenance strategies to optimize this trade-off.

Cost relationships in LoRa-enabled TES demonstrate particularly interesting cascading effects. The technology's long-range capabilities reduce infrastructure requirements through a saving rate mechanism (R4) that interacts with adaptive power control features (R3) to create compound cost benefits. These benefits manifest through reduced operational expenses and decreased infrastructure needs. However, our analysis reveals a sophisticated balancing mechanism (B5) where implementation costs moderate these advantages through multiple pathways involving infrastructure and manufacturing impacts. This interaction suggests that implementation timing and scaling strategies significantly influence overall cost effectiveness.

The security and reliability aspects of the system reveal previously unexplored connections. While LoRa's encrypted communication and device-level authentication create strong initial security benefits (B1, B2), these mechanisms interact with signal penetration capabilities (R10) through frequency synthesizer employment. This interaction suggests that security measures might have unintended consequences on system reliability - a critical consideration for system designers. The reliability analysis further reveals how Chirp Spread Spectrum modulation supports uninterrupted communication, though increasing device density introduces potential interference that must be carefully managed.

Environmental feedback loops demonstrate more nuanced benefits than initially apparent. The reduction in e-waste (R7) connects directly to system efficiency through extended battery lifespan, while system-wide communication improvements (R8) influence manufacturing impacts. Perhaps most significantly, the analysis reveals how weather forecasting capabilities create a direct link to renewable energy system utilization (R9), suggesting environmental benefits might accumulate more rapidly than previously understood. This finding has important implications for long-term sustainability planning.

The multiple pathways affecting system costs and performance indicate the need for careful phasing of technology adoption. Critical intersection points emerge where non-essential data transmission interacts with adaptive power control, suggesting opportunities for optimizing communication strategies. Similarly, the relationship between frequency synthesizer employment and system reliability reveals potential bottlenecks that must be considered in scaling plans.

These interconnected feedback mechanisms suggest several important considerations for LoRa integration in TES. Implementation planning must account for both immediate benefits and longer-term scaling effects, particularly where security measures interact with system reliability. System designers should anticipate and plan for transition points where balancing loops begin to counteract initial benefits,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

while recognizing that environmental benefits may continue to accumulate through multiple reinforcing loops with fewer balancing constraints.

This systems analysis provides valuable insights for future TES development while highlighting areas requiring further investigation. Particularly important is the need to understand how these feedback mechanisms behave under different deployment scales and environmental conditions, especially considering the complex interactions between security, reliability, and system performance.

V. Empirical Validation of Systems Dynamics Model

System dynamics models are designed to capture causal mechanisms and feedback processes, but their credibility depends on showing that these theoretical dynamics correspond to observable system behavior. Our causal model includes sixteen feedback loops spanning technical, economic, and environmental dimensions. A full validation of all loops is beyond the scope of this study, not least because many—such as security/adoption dynamics, cost trade-offs, and environmental impacts—require field experiments or lifecycle analyses that cannot be addressed through simulation alone.

Instead, this study concentrates on the technical core of the model: the feedbacks that govern LoRa's role as a communication layer within TES. These loops are both central to system performance and empirically testable using simulation tools, making them the most effective entry point for validation. By establishing that the physical-layer dynamics behave as expected, we provide a foundation for future work on higher-level socio-economic and environmental interactions.

Our analysis focuses on six loops, organized into two themes that capture the key trade-offs in LoRa-enabled TES:

1) **Energy Efficiency & Operational Viability**: This theme highlights reinforcing mechanisms that make LoRa an attractive option for TES. It encompasses R1 (Energy Consumption Reduction) and R3 (Adaptive Power Control Optimization). Our central hypothesis is that LoRa's adaptive features will cause significant, quantifiable reductions in network-wide energy consumption, validating the powerful reinforcing nature of these loops.
2) **Scalability & Network Reliability**: This theme examines the tension between system expansion and operational limits. It incorporates R6 (DER Monitoring Effectiveness) and R10 (System Reliability) alongside the balancing loop B3 (Network Congestion). Here, the hypothesis is that adaptive configurations improve communication reliability (R6, R10), but only up to a saturation point where congestion effects dominate (B3).

Together, these six loops form the technical backbone of the causal model. They determine whether LoRa can serve as a viable communication layer for TES and can be empirically validated using network simulation. Loops concerning economic trade-offs, environmental impacts, and adoption dynamics remain theoretically specified but are reserved for validation through empirical fieldwork in future studies.

A. Experimental Design

To test the hypotheses derived from the causal loop model, we employed LoRaSim [29], [30], an open-source packet-level simulator widely used to study the scalability, energy efficiency, and reliability of LoRa networks. To capture deployment diversity, the number of participating nodes was scaled from 50 (representative of small microgrid installations) to 1,500 (reflecting larger community-level systems). Each node was modeled as a LoRa-enabled device, such as a sensor or distributed energy resource controller. Simulation runs lasted 1,000,000 ms (approx. 16.7 minutes), a horizon long enough to ensure multiple packet transmissions per node despite LoRa's low duty cycle. Within this framework, we implemented three canonical configurations from the LoRaSim codebase—labeled Experiment 0, Experiment 3, and Experiment 5—that progressively introduce adaptive mechanisms:

- **Condition 1 (Baseline Case - Exp. 0)**: A non-adaptive network with fixed spreading factor (SF12) and fixed transmit power. This case isolates the effects of congestion dynamics (loop B3) by excluding adaptive responses.
- **Condition 2 (Partially Adaptive Case - Exp. 3)**: Adaptive spreading factor with fixed transmission power. This configuration highlights how spectral efficiency improvements influence delivery reliability, directly probing loops R6 (DER Monitoring Effectiveness) and R10 (System Reliability).
- **Condition 3 (Fully Adaptive Case - Exp. 5)**: Joint adaptation of spreading factor and transmit power. This condition represents the optimized system state, where the reinforcing effects of energy efficiency and reliability dynamics operate in tandem.

For each run, we collected five key performance indicators: (i) total packets transmitted, (ii) successful deliveries, (iii) collisions, (iv) packet delivery ratio (PDR), and (v) total energy consumption across nodes. Together, these metrics capture the fundamental trade-off between reliability and efficiency. Finally, the radio environment was standardized across experiments with a carrier frequency of 860 MHz, bandwidth of 125 kHz, and payload size of 20 bytes. These parameters mirror common LoRa configurations reported in prior studies [29], [30], ensuring both comparability and practical relevance.

B. Validation Results

The simulations yielded clear, quantitative evidence that supports the hypotheses and validates the behavior of the core feedback loops. Two sets of dynamics are especially prominent: the reinforcing effects driving energy efficiency, and the tradeoff between network reliability and congestion.

1) Energy Efficiency Reinforcing Loops (R1, R3): The results in Figure 4 show that adaptive configurations deliver dramatic improvements in energy efficiency, consistent with the reinforcing loops R1 and R3. At 50 nodes, total energy use dropped from 937.22 J in the baseline case (blue) to

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

11.84 J with adaptive spreading factor (orange) and 8.29 J with full adaptation (green). The greater than 99% reduction observed in the fully adaptive case demonstrates the powerful compounding effect of these loops. Moreover, the additional 30% efficiency gain achieved by moving from spreading-factor adaptation (Condition 2) to joint spreading and power adaptation (Condition 3) isolates the specific contribution of adaptive power control, providing direct empirical validation of loop R3.

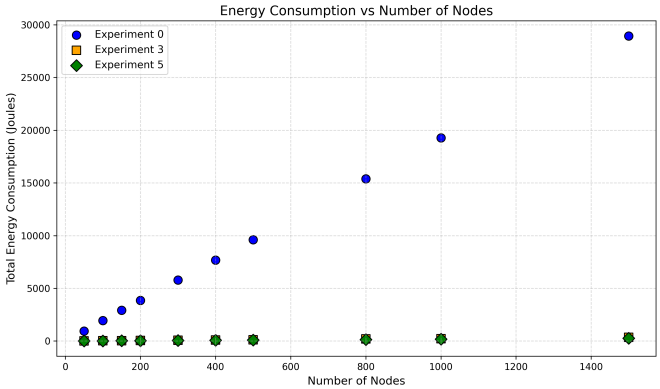


Figure 4. Energy consumption as a function of the number of nodes for Experiments 0, 3, and 5

2) Reliability–Congestion Tradeoff (R6, R10 vs. B3): Adaptive configurations also validated the expected tension between reliability and congestion (Figure 5, Table I). At small scales, LoRa's adaptive features sustained high packet delivery ratios (PDRs), essential for continuous DER monitoring (R6). With 50 nodes, adaptive cases achieved PDRs above 90%, while the baseline collapsed to just 0.14%. This resilience under increasing load confirms the reliability-enhancing role of adaptive mechanisms (R10).

However, the results also reveal the counteracting influence of the congestion loop (B3). In the baseline, the network collapsed completely beyond 100–200 nodes, with collision rates exceeding 99%. Even in adaptive scenarios, performance began to plateau around 800–1000 nodes, with congestion ultimately eroding reliability. These results confirm B3 as a fundamental scalability limit: adaptation delays but does not eliminate the onset of congestion-driven performance decline.

TABLE I. Key Simulation Results at 50 Nodes

Experiment	Packets Sent	Collisions	PDR (%)	Energy (J)
Experiment 0	4147	4141	0.14	937.22
Experiment 3	5002	388	92.0	11.84
Experiment 5	4900	410	91.6	8.29

The simulations provide clear empirical support for the technical core of our system dynamics model. We validated

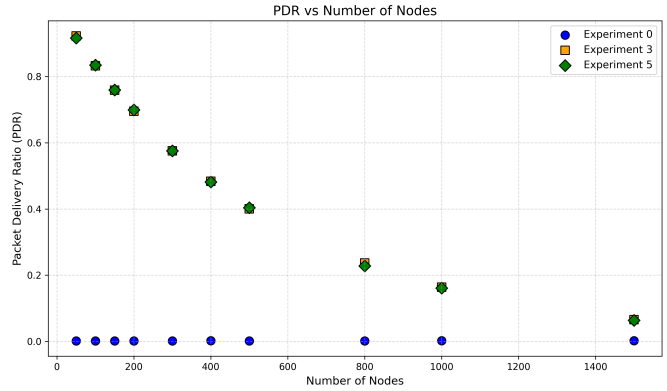


Figure 5. Packet delivery ratio as a function of the number of nodes for Experiments 0, 3, and 5

two central mechanisms: (i) the reinforcing effects of adaptive communication on energy efficiency and reliability, and (ii) the balancing influence of congestion in constraining scalability. Together, these findings confirm that the physical-layer dynamics embedded in loops R1, R3, R6, R10, and B3 behave as theorized.

Grounding these foundational loops in data strengthens the credibility of the broader model. Higher-level socio-economic and environmental mechanisms (e.g., cost trade-offs in R4 and B5, environmental impacts in R7) rely on the assumption that LoRa can provide scalable, efficient communication. With that assumption now empirically established, those loops can be pursued in future studies using methods such as techno-economic modeling, lifecycle analysis, or behavioral research.

For system design, the results indicate two key implications. First, adaptive LoRa should be treated as the baseline configuration for TES: fixed-parameter networks collapse at even modest scales and are therefore untenable in practice. Second, while adaptation substantially extends network viability, the observed performance ceiling at 800–1000 nodes highlights the need for architectural strategies—such as segmentation, clustering, or hybrid communication layers—when scaling beyond community-level deployments.

Table II summarizes the validation status of the full causal model, clarifying both the technical scope achieved here and the remaining loops reserved for future empirical investigation.

VI. Limitations

This study's findings are shaped by several assumptions and technical constraints that must be acknowledged. The causal loop analysis assumes relatively stable interactions between system components and rational stakeholder behavior. In practice, energy markets exhibit nonlinear dynamics and bounded rationality, which could alter outcomes. The study also evaluates LoRa based on its current documented capabilities, though future protocol or hardware advancements may shift its role in TES communication.

From a technical perspective, LoRa's low data rates (22 bps to 27 kbps) and asymmetric uplink/downlink design constrain

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE II. Mapping of Feedback Loops to Simulation Validation

Loop	Description	Simulation Test	Input Parameters	Expected Outcome	Actual Result	Status
R1	Energy consumption reduction	Exp 0, 3, 5	Fixed SF/Power vs Adaptive SF/Power	Lower energy with adaptive configuration	Exp 0: 937.22 J → Exp 3: 11.84 J → Exp 5: 8.29 J	Validated
R3	Adaptive power control	Exp 3 vs 5	With vs Without Power Control	Further energy reduction	Exp 3: 11.84 J → Exp 5: 8.29 J	Validated
R6	DER monitoring effectiveness	All experiments	PDR across node densities	Higher PDR with adaptive configuration	Exp 0 PDR: 0.14% → Exp 3: 92% → Exp 5: 91.6%	Validated
R10	System reliability	All experiments	PDR vs Node Density	Reliability maintained under higher load	PDR sustained longer in Exp 3 & 5	Validated
B3	Network congestion	Exp 0	High collision rates with increasing nodes	PDR degradation at higher node counts	PDR collapsed after ∼100–200 nodes	Validated
B4	Operational burden	Not modeled	N/A	Maintenance limits scalability	Not simulated – LoRaSim lacks battery replacement model	Not validated
B1	Security vs adoption	Not modeled	N/A	Security saturation effects	Not simulated – LoRaSim lacks security/authentication modules	Not validated
R4, R7–R9, B5	Various (cost, e-waste, forecasting)	Not modeled	N/A	Economic/environmental benefits	Not measured in current setup	Not validated

its suitability for real-time TES operations. While LoRa excels in coverage and energy efficiency, its limited throughput makes it less capable of supporting functions such as dynamic price signaling, rapid control actions, or emergency interventions. The results therefore suggest LoRa is best positioned as part of a hybrid communication architecture rather than a standalone TES backbone.

The simulated environment excluded several real-world factors that could affect performance. Physical effects such as attenuation, multipath interference, or weather conditions were not modeled, nor was coexistence with other technologies (e.g., Wi-Fi, LTE, or co-located LoRa networks) [29]. Node mobility and asynchronous traffic patterns were also omitted, as all nodes transmitted at fixed intervals. Energy consumption estimates relied on simplified transmission models rather than detailed discharge profiles. Adaptive power control and spreading factor assignment were implemented using approximated link-quality metrics (RSSI, SNR) rather than dynamically computed values. These simplifications likely overstate the stability and efficiency of the simulated system.

Validation was restricted to communication-layer performance—packet delivery, collisions, and energy use. Higher-level feedback loops identified in the system dynamics model remain unvalidated. These limitations indicate that while the study provides robust insights into LoRa's role in TES, it does not capture the full complexity of real-world deployments. The results should be interpreted as establishing the technical feasibility and constraints of LoRa-based TES communication under idealized conditions, rather than as a definitive performance benchmark.

VII. Conclusion and Future Directions

This systems dynamics analysis of LoRa integration into TES reveals how communication technology choices shape performance through intertwined reinforcing and balancing feedback loops. While LoRa's efficiency and cost advantages encourage adoption, scaling introduces congestion and throughput constraints that limit its effectiveness. These dynamics suggest that TES performance is governed not by linear improvements but by scale-dependent trade-offs.

Simulation experiments reinforced this perspective. Adaptive configurations—particularly those combining spreading factor adjustment and power control—delivered dramatic improvements in energy efficiency and reliability, validating the reinforcing loops predicted in the model. Yet, as node density rose, congestion effects eroded these gains, highlighting a scalability ceiling and confirming the balancing mechanisms. LoRa thus appears well suited for small to medium deployments but insufficient alone for larger systems, where hybrid architectures with higher-bandwidth technologies are required.

These findings carry implications for both design and policy. System architects must account for long-term scaling effects rather than relying solely on early efficiency gains, and researchers should investigate hybrid communication models to extend LoRa's utility. Future work should also refine the tipping points between reinforcing and balancing dynamics through quantitative modeling, field validation, and exploration of institutional factors such as regulation and market design. More broadly, this study demonstrates the value of systems thinking in energy technology evaluation, linking communication-layer performance to the socio-economic and environmental dimensions of sustainable energy transitions.

References

[1] A. Zarei, E. Honu, M. Mansouri, and P. Odonkor, "A systems dynamics analysis of communication technology integration in complex transactive energy systems," in *Proc. ENERGY 2025: The Fifteenth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*, 2025, pp. 68–74.

[2] M. A. Tabar, M. A. Jirdehi, and M. Shaterabadi, "Impact of bi-facial pv panels' presence as the novel option on the energy management and scheduling of the interconnected grids: Comprehensive outlook," *Journal of Building Engineering*, vol. 90, p. 109495, 2024.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[3] P. Koukaras, K. D. Afentoulis, P. A. Gkaidatzis, A. Mystakidis, D. Ioannidis, S. I. Vagropoulos, and C. Tjortjis, "Integrating blockchain in smart grids for enhanced demand response: Challenges, strategies, and future directions," *Energies*, vol. 17, no. 5, p. 1007, 2024.

[4] F. Lezama, J. Soares, P. Hernandez-Leal, M. Kaisers, T. Pinto, and Z. Vale, "Local energy markets: Paving the path toward fully transactive energy systems," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 4081–4088, 2019.

[5] M. Daneshvar, M. Pesaran, and B. Mohammadi-Ivatloo, "Transactive energy in future smart homes," in *The Energy Internet*, W. Su and A. Q. Huang, Eds. Woodhead Publishing, 2019, pp. 153–179.

[6] Z. Liu, L. Wang, and L. Ma, "A transactive energy framework for coordinated energy management of networked microgrids with distributionally robust optimization," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 395–404, 2020.

[7] B. H. Rao and P. S. M, "Prosumer participation in a transactive energy marketplace: A game-theoretic approach," in *2020 IEEE International Power and Renewable Energy Conference*, Oct. 2020, pp. 1–6.

[8] S. Zhang, D. May, M. Gül, and P. Musilek, "Reinforcement learning-driven local transactive energy market for distributed energy resources," *Energy and AI*, vol. 8, p. 100150, 2022.

[9] D. Xu, B. Zhou, N. Liu, Q. Wu, N. Voropai, C. Li, and E. Barakhtenko, "Peer-to-peer multienergy and communication resource trading for interconnected microgrids," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2522–2533, 2021.

[10] P. Siano, G. D. Marco, A. Rolán, and V. Loia, "A survey and evaluation of the potentials of distributed ledger technology for peer-to-peer transactive energy exchanges in local energy markets," *IEEE Systems Journal*, vol. 13, no. 3, pp. 3454–3466, 2019.

[11] O. Abrishambaf, F. Lezama, P. Faria, and Z. Vale, "Towards transactive energy systems: An analysis on current trends," *Energy Strategy Reviews*, vol. 26, p. 100418, 2019.

[12] Z. Sun, H. Yang, K. Liu, Z. Yin, Z. Li, and W. Xu, "Recent advances in lora: A comprehensive survey," *ACM Transactions on Sensor Networks*, vol. 18, no. 4, 2022.

[13] C. Li and Z. Cao, "Lora networking techniques for large-scale and long-term iot: A down-to-top survey," *ACM Computing Surveys*, vol. 55, no. 3, 2022.

[14] O. Georgiou and U. Raza, "Low power wide area network analysis: Can lora scale?" *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 162–165, 2017.

[15] O. O. Tooki and O. M. Popoola, "A comprehensive review on recent advances in transactive energy system: Concepts, models, metrics, technologies, challenges, policies and future," *Renewable Energy Focus*, vol. 50, p. 100596, 2024.

[16] N. Dhirasasna and O. Sahin, "A system dynamics model for renewable energy technology adoption of the hotel sector," *Renewable Energy*, vol. 163, pp. 1994–2007, 2021.

[17] R. G. Newell and D. Raimi, "Global energy outlook comparison methods: 2020 update," pp. 1–27, 2020.

[18] Y. Sugiawan and S. Managi, "New evidence of energy-growth nexus from inclusive wealth," *Renewable and Sustainable Energy Reviews*, vol. 103, pp. 40–48, 2019.

[19] M. Mallaki, M. S. Naderi, M. Abedi, S. D. Manshadi, and G. B. Gharehpetian, "A novel energy-reliability market framework for participation of microgrids in transactive energy system," *International Journal of Electrical Power and Energy Systems*, vol. 122, p. 106193, 2020.

[20] K. Polat and A. Ozdemir, "Impacts of transactive energy trading on the load point reliability indices of lv distribution system," *IEEE Access*, vol. 11, pp. 132 119–132 130, 2023.

[21] M. Jalali, K. Zare, and S. Tohidi, "Designing a transactive framework for future distribution systems," *IEEE Systems Journal*, vol. 15, no. 3, pp. 4221–4229, 2021.

[22] H. Zhou, B. Li, X. Zong, and D. Chen, "Transactive energy system: Concept, configuration, and mechanism," *Frontiers in Energy Research*, vol. 10, 2023.

[23] X. Ge, B. Yang, J. Ye, G. Mao, C. X. Wang, and T. Han, "Spatial spectrum and energy efficiency of random cellular networks," *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 1019–1030, 2015.

[24] L. Fu, X. Fu, Z. Zhang, Z. Xu, X. Wu, X. Wang, and S. Lu, "Joint optimization of multicast energy in delay-constrained mobile wireless networks," *IEEE/ACM Transactions on Networking*, vol. 26, no. 1, pp. 633–646, 2018.

[25] J. P. S. Sundaram, W. Du, and Z. Zhao, "A survey on lora networking: Research problems, current solutions, and open issues," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 1, pp. 371–388, 2020.

[26] S. Devalal and A. Karthikeyan, "LoRa technology—an overview," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 1–6.

[27] X.-C. Le, B. Vrigneau, M. Gautier, M. Mabon, and O. Berder, "Energy/reliability trade-off of LoRa communications over fading channels," in *International Conference on Telecommunications*, Saint-Malo, France, Jun 2018, hAL Id: hal-01816574.

[28] S. S. Borkotoky, J. F. Schmidt, U. Schilcher, P. Battula, and S. Rathi, "Reliability and energy consumption of lora with bidirectional traffic," *IEEE Communications Letters*, vol. 25, no. 11, pp. 3743–3747, 2021.

[29] M. Bor, U. Roedig, T. Voigt, and J. M. Alonso, "Do lora low-power wide-area networks scale?" in *MSWiM 2016 - Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2016, pp. 59–67.

[30] R. Marini, K. Mikhaylov, G. Pasolini, and C. Buratti, "Lorawansim: A flexible simulator for lorawan networks," *Sensors*, vol. 21, no. 3, pp. 1–19, 2021.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Emulating the Style of Johannes Brahms: Neural Network-Based Generation of Romantic Piano Music

James Doherty
Technological University Dublin
Central Quad, Grangegorman Lower, Dublin 7, D07ADY7
Dublin, Ireland
e-mail: j_doherty1992@hotmail.com

Brendan Tierney
Technological University Dublin
Central Quad, Grangegorman Lower, Dublin 7, D07ADY7
Dublin, Ireland
e-mail: brendan.tierney@tudublin.ie

***Abstract* – Neural network architectures currently are only able to employ music generation tasks to similar levels of human composers when the music is at a basic compositional standard. Research has shown that various neural network models struggle with the complex motifs and harmonic structures of Western Classical Music. This study aims to determine if various data preprocessing and augmentation techniques can train a neural network model to generate pieces of piano music to a similar level of musicality and emotion as Romantic Period composer Johannes Brahms. Quantitative experimentation involving Music Information Retrieval was conducted as well as a quantitative survey with respondents consisting of only professional musicians, composers and conductors. Analysis of the results from the quantitative experiment demonstrated that the Transformer models using Self-Attention, Relative Self-Attention and Local Windowed Attention generated statistically indistinguishable results to the original piano works of Brahms in terms of duration distribution, pitch class distribution, rhythmic variation and pulse clarity. Analysis of the quantitative survey discovered that participants struggled to distinguish between the pieces generated by Brahms and the models, with two generated pieces being at majority mistaken as one of Brahms' own works. The results indicate that various data preprocessing and augmentation methods do have an impact on model accuracy and for numerous musical characteristics, can be statistically indistinguishable to Brahms through quantitative experiment and survey. Further research is needed to identify other techniques to improve the musical characteristics of entropy and global energy in the generated pieces so that they are at a statistically comparable level to Brahms. The paper discovers that transformer models using varying attention mechanisms were able to generate longer sequences of music up to 2 minutes long which contained the composite motifs and harmonic structures of romantic period piano music.**

Keywords-artificial Intelligence; music generation; neural network architecture; Brahms.

I. INTRODUCTION

The intention of this project is to generate piano music in the style of classical music composer Johannes Brahms by training a Recurrent Neural Network (RNN), a Long Short-Term Memory (LSTM) based RNN, various Transformer models with different attention mechanisms, and a Perceiver AR model [1]. These models will be trained with a pre-processed and augmented dataset containing MIDI files of Brahms' piano works. To deem the success of the project, the best musical pieces generated from the neural network models must show statistical indistinguishability from the Brahms corpus in various musical variables using Music Information Retrieval (MIR). Along with this, the pieces must also be mistaken by professional musicians, composers, and conductors as one of Brahms' own works through a quantitative survey. Although there have been many examples of AI models generating music in the style of particular composers, no models have been created to generate the work of Brahms. The lack of a detailed computational analysis of Brahms shows a gap in the study of romantic period composers, which Brahms was a key figure of [2]. According to studies taken, computer-generated music has traditionally only sounded human-like when short excerpts were created and struggled with the complex motifs and harmonic cadences of romantic period piano music. This could be down to them being poor at handling higher-level musical structures due to the models only learning how to play the next note according to the previous. Zheng et al. (2022) stated that their model was limited due to it learning how to play the next note according to the previous ones however it did not have any knowledge on the structure of music [3]. In their paper, Child et al. developed a sparse transformer and stated that it was able to extract complex patterns from sequences up to 30 times longer than possible previously [4]. Likewise, Hawthrone et al. developed a Perceiver AR model which had the ability to efferently handle longer sequences with an improved memory efficiency [5]. After listening to the

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

examples from the papers, the generated pieces still consisted of basic harmonic and rhythmic structures and struggled with the complex motifs and harmonic cadences of romantic period piano music. The most common quantitative experiment was surveys. It appears that simple questions were asked to a small group of test subjects who were mainly computer science students and participants who had little to no music compositional knowledge. A more realistic survey would involve professional musicians, composers and conductors who specialise in the subject and would give a critical and educated opinion on the generated pieces compared to the composers that they have great knowledge about.

The importance of the research problem not only addresses AI's ability to generate music, but also highlights potential significance on how music could be composed in the future, particularly for those with no previous musical knowledge [6]. The research assumes that neural network models can already be trained to learn the general characteristics and patterns of various musical composers. To help with producing optimal results, the selected MIDI files for the dataset were exact replications of Brahms' piano music without errors and inconsistencies.

Based on the issues raised above, this study focused on understanding what configurations of the neural network models performed best when tasked with producing music in the style of Brahms. Therefore, the research question is:

To what extent can the accuracy of various Neural Network Models, trained with Long Short-Term Memory and numerous Attention mechanisms, be significantly improved by augmenting MIDI files containing the compositional works of Johannes Brahms with an augmentation pipeline to generate pieces of music that are mistaken by professional musicians, composers and conductors as one of Brahms' own works?

Due to conclusions from studies taken and the general state of knowledge at the time of beginning the project, the null hypothesis for this research project is;

H0: Neural network models cannot generate piano works to the same level of musicality and emotion as Brahms. Due to this, generated pieces will not be statistically indistinguishable through music information retrieval or mistaken as a work of Brahms by professional musicians, composers and conductors through a quantitative survey using Likert Scales.

Through the utilisation of an augmentation pipeline to expand the MIDI dataset containing the compositional piano works of Johannes Brahms, more musical variations could be created including transposition, rhythmical and note durations. In addition to this, various preprocessing techniques including track splitting, quantisation and normalisation could help make the MIDI files more readable for the models. This provides an alternate hypothesis;

H1: If an augmentation pipeline is utilised to expand a MIDI dataset of pre-processed files containing the piano works of Johannes Brahms, then various neural network models trained with Long Short-Term Memory and numerous Attention mechanisms could generate pieces of music that is statistically indistinguishable from Brahms and could be mistaken as one of Brahms' own piano works by professional musicians, composers and conductors through a quantitative survey using Likert Scales and various Independent-Samples T-Tests and Hotelling's T^2 Tests being implemented to determine whether the p-value is > 0.05 to support the conclusion of statistical indistinguishability

This paper contains a total of four sections. Section II describes the experiment design, methodology and how the dataset was prepared and pre-processed. Along with this, the neural network models obtained for the project and MIR functions will be explained as well as ethical considerations. Section III analyses and evaluates the results of the quantitative experiment and survey to determine if the experiments provide evidence that the null hypothesis is incorrect. Section IV summarises what has been learnt and proposes recommendations and adjustments for future studies.

II. DESIGN AND METHODOLOGY

The research project was carried out through three stages. Firstly, data was collected, pre-processed, and augmented. The dataset contained 67 MIDI files of Johannes Brahms' piano works and was obtained for offline manipulation from *Classical Archives* and *MIDIworld*. MIDI (Musical Instrument Digital Interface) is seen as one of the most important tools for both musicians and producers as it represents a method to store and transmit musical data. They contain *Note-on* and *Note-off* messages to indicate the beginning and end of each note and delivers information about various musical characteristics in musical notes, including pitch, duration and dynamics.

An augmentation pipeline was employed to create variations in melodies, tempo, rhythm, and transpositions. This was followed by obtaining and training various neural network models with the augmented dataset. The models were analysed for training accuracy and loss to determine the best configurations. After training, the best generated examples from each model were analysed through various MIR functions using *MIDIToolbox* and *MIRToolbox* to determine the best performing models [7][8]. Finally, the best models generated pieces of music that were evaluated against the pieces of Brahms' repertoire for statistical equivalence. Along with this, the same generated pieces were used in a quantitative listener survey to test professional musicians, composers and conductors on whether they could differentiate between the generated pieces against the Brahms original. All quantitative experiments involved evaluation to test the research hypothesis through Independent-Samples T-Tests and Hotelling's T^2 Tests.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A. Preparation and Preprocessing of Dataset

The preparation and preprocessing of the dataset involved numerous steps to optimise the potential of training the neural network models. Track splitting involved dividing the MIDI tracks into smaller segments of 30 second clips to make the tracks more digestible for the models in training [9]. The conversion of multi-track MIDI files into a single monophonic track allowed for further simplicity in the files. Normalisation was performed to ensure that note velocities of the files were standard to provide consistent values for training and evaluation tasks. Finally, all the MIDI files were quantized to semiquavers to adjust the timings of notes and align them with the correct timing to ensure consistency in rhythm and therefore making it easier for the models to digest them [10]. Figure 1 displays an example of Brahms' *Variations on a Hungarian Song in D major, Op.21, No. 11* being quantised to semiquavers, also known as sixteenth notes.

Depending on the model being used, there were several packages utilised to extract information from MIDI files:

- **Pyfluidsynth:** Python package that enables audio playback of MIDI files in a Google Colab setting [11].
- **pretty_midi:** Package that contains various functions and classes for manipulating MIDI data so that it could be easily modified for information extraction. Figure 2 shows an example of pretty_midi extracting variables from each note [12].

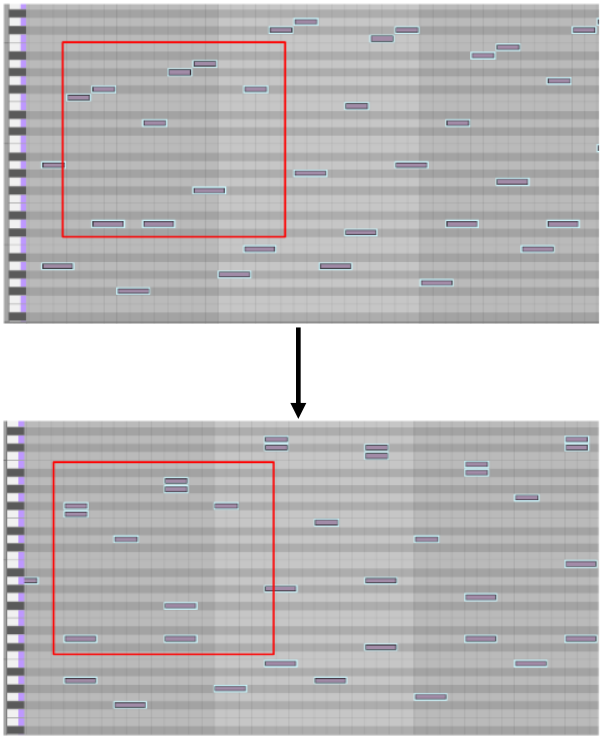


Figure 1. Example of Sixteenth note Quantisation

```
for i, note in enumerate(instrument.notes[:10]):
  note_name = pretty_midi.note_number_to_name(note.pitch)
  duration = note.end - note.start
  print(f'{i}: pitch={note.pitch}, note_name={note_name},'
        f' duration={duration:.4f}')
```

```
0: pitch=79, note_name=G5, duration=0.2487
1: pitch=67, note_name=G4, duration=0.2708
2: pitch=67, note_name=G4, duration=0.1589
3: pitch=79, note_name=G5, duration=0.3060
4: pitch=67, note_name=G4, duration=0.2201
5: pitch=79, note_name=G5, duration=0.3346
6: pitch=64, note_name=E4, duration=0.0977
7: pitch=67, note_name=G4, duration=0.2943
8: pitch=79, note_name=G5, duration=0.3958
9: pitch=72, note_name=C5, duration=0.2005
```

Figure 2. Example of *pretty_midi* extracting variables from each note

- **NoteSequences:** Numerical representations of music notes including pitch number representation, along with note start and end time [13].

An augmentation pipeline was utilised to create variations in melodies, tempo, rhythm and transpositions. Several techniques were used to increase the dataset size. Time-stretching was applied to make each MIDI file 5% faster or slower. Another method was to transpose each of the MIDI files so that the pitches would be raised or lowered by a third [9]. Doing these increased the dataset by 500% with a total of 445 tracks. The augmented dataset was divided into training and testing sets which were defined at an 90%/10% split with repeated Montecarlo sampling (100 times). This same split was used for all the models to conduct a fair experiment. In comparison with prior studies, data preparation and preprocessing was influenced by previous research, which was collected, adapted and integrated into this paper.

B. Neural Network Models

Numerous Neural Network Architectures were obtained for offline manipulation and trained with the augmented Brahms MIDI dataset. A Recurrent Neural Network was acquired from TensorFlow [12]. A RNN was described by IBM as "a type of artificial neural network which used sequential data or time series data" [14]. Sequential data was utilised to predict the next output based off previous elements in the sequence. For this model, the training dataset was created through extraction of notes from the MIDI files with three values being taken from each note; pitch, step and duration. The model was trained on batches of note sequences with each example consisting of a sequence of notes as the input tokens and the subsequent notes served as the target output. This meant that the model was trained to predict the next note in the sequence containing 100 notes as the input, a common practice used in text classification. Similar to the training dataset, the model contained pitch, step and duration outputs. For the step and duration outputs, a loss function was constructed based on *mean squared error* equation to encourage the

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

model to output non-negative values. Equation (1) shows the formula were y_i represents the nth token value and $\hat{y}_i$ being the predicted value from the model. n represented the number of observations [15].

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \quad (1)$$

Evaluating the model showed that *pitch_loss* was much greater that *duration_loss* and *step_loss.* This result indicated that predicting the next note had proven to be a more difficult task when compared to the other two. The model was trained with 22 epochs and testing recorded a training loss of 0.550 and accuracy of 0.628. RNNs suffer from vanishing gradient problem. With the neural network using the gradient decent algorithm to update the weight, the gradients therefore decreased in growth the further down the layers the network progressed.

A solution to this problem is the use of LSTM which utilises gating mechanisms to control the movement of information and gradients to allow for the network to learn and maintain information over longer sequences. [16] An LSTM-based RNN was obtained from Huang et al. [17]. The RNN utilised event sequence encoding with the following representing the event types:

- **NOTE_ON:** The start of a musical note
- **NOTE_OFF:** The end of a musical note
- **TIME_SHIFT:** Change of time (bpm)
- **VELOCITY:** Change of attack on notes

The model supported expressive timing with *time_shift* events allowing it to control the tempo at increments between 10ms and 1 second. *Velocity* events allowed for a dynamic to be set at the current *note_on* event which helped build more natural tension and timing. For the model to be able to read the Brahms MIDI dataset, they were converted to *NoteSequences*, a numerical representation of musical notes including pitch number representation, along with the note start and end times [13]. The model was trained for a total of 10002 checkpoints accumulating to 17 hours of training. Although LSTM provides a solution to vanishing gradient problem. They still suffer from gradient explosion which occurs when gradients become too large, causing a model to become unstable and unable to learn from training data. LSTMs struggle to memorise earlier information when very long sequences were being processed due to its limited context window. As the forget gate became more prominent, older information was replaced by new data.

Various Transformer models containing different attention mechanisms were obtained from Project Los Angeles [18]. Proposed by Google researchers Vaswani et al., transformer models do not rely on recurring processing of data. Instead, they operate on an attention mechanism [19]. Attention allows for neural networks to concentrate on particular parts of the input. Transformer models contain two main components, an encoder, which produces a sequence of hidden states from an input, and a decoder, which takes the hidden states from the encoder and generates an output. The attention mechanism enables the decoder to access the encoder's hidden states meaning that the decoder can learn long-range dependencies in the input. This feature allows for the models to effectively handle larger amounts of data [20]. When compared to Simon & Oore's (2017) *Performance RNN* model, Huang et al. (2018) stated that while the LSTM-based model was able to generate credible music for very short samples, there was an apparent lack of long-term structure. However, their model *Music Transformer* could generate music with a consistent style that created multiple phrases out of a single motif. The attention mechanisms obtained for the transformer models tested were:

- **Self-Attention** – Processes inputs in the same sequence, enabling the model to capture dependencies within the input [21].
- **Relative Attention** – The relative position of tokens is considered based on similarity of other tokens in the sequence [22].
- **Local Windowed Attention** – Restricts attention to a fixed window of tokens, enabling the model to focus on nearby information [23].
- **Relative Self-Attention** – Combination of Self and Relative attention allowing the model to focus on relevant information based on the positional relationships of tokens [17].
- **Sparse Attention** – Attends to a subset of tokens instead of the entire sequence to improve efficiency while retaining information [24].

Traditional transformers with self-attention grow quadratically with the sequence length. This form of scaling could restrict the model's ability to work with longer sequences due to it requiring lots of memory and computational power. Sparse attention addresses this issue with the introduction of sparse factorisations of the attention matrix and therefore attending to a reduced subset of elements strategically chosen by the model. This enables the mechanism to still capture necessary information whilst the model performance is greatly increased [24]. Equation (2) displays the reformulation of the self-attention mechanism for sparse attention.

$$O(n^2) \longrightarrow O(n\sqrt{n}) \quad (2)$$

To train the transformer models, the MIDI files had to be processed with the TMIDIX MIDI processor provided by Tegridy code which converted the MIDI files to a score representation so that it could be understood by the model [25].

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Pieces were generated in the style of OpenAI's *MuseNet* workflow where a premier seed MIDI file would be input and treated as the beginning of the generated piece. Doing so gave information to the model of the tempo, style and dynamics of the piece. A continuation generator was then implemented to append a chosen number of tokens and the multiple generations were batched to append the best results. This provided a degree of outcomes and the model was tested with variations on temperature, generation tokens and memory tokens. According to Huang & Yang (2020), amongst the great number of deep learning-based models that had been proposed for automatic music composition, the Transformer "stood out as the prominent approach for generating expressive classical piano performance with a coherent structure of up to one minute" [26].

A Perceiver Autoregressive (AR) model was also obtained from Project Los Angeles. Seen as an improvement to the Transformer model using latent array to distinguish the size of inputs and outputs, allowed for the model to efferently handle longer sequences with an improved memory efficiency [27]. Perceiver AR expands from the foundations of the original Perceiver Model with the addition of autoregressive capabilities, which was the ability to generate outputs sequentially based off previously generated outputs. This, along with the use of cross-attention to encode inputs into latent space, gave it the ability to train up to 100,000 elements. Perceiver AR contains 1024 latents and 24 self-attention layers. Due to each new note being predicted based off the preceding full sequence, the model can generate pieces of music containing high levels of melodic, harmonic and rhythmic consistency. The processing of MIDI files and generation of music was done so in a similar style to that of the transformer models.

C. *Quantitative Experiment & Survey*

A quantitative experiment was conducted to evaluate and compare the generated pieces of music from each of the models using Music Information Retrieval (MIR). MIR is the extraction of significant audio features from music data to be analysed through various algorithms and techniques. Raw audio signals were extracted and analysed from Waveform Audio Files (.wav) including pitch, harmony, rhythm and timbre that acquired an expressive description that is machine processible. MIR was used to test the neural network model's ability to replicate the musical characteristics and motifs of Brahms' piano works. Various MIR variables were utilised to gather various quantitative data from the models to test against not only each other, but the original works of Brahms. These variables included:

- **Entropy** – The measure of uncertainty or unpredictability
- **Duration Distribution** – The statistical analysis of note durations as well as silence
- **Pitch Class Distribution** – The evaluation of frequency of musical pitches
- **Mean Roughness** – The measure of dissonance or clashing sounds
- **Global Energy** – The total amount of energy held within a waveform. Calculated using the Root Mean Square (RMS) (3), each audio signal sample's amplitude x_i was squared with the mean of the square values being calculated to which the square root of that mean is taken. N represented the total number of samples in the signal [28].
- **Normalised Pairwise Variability Index (nPVI)** – The analysis of variability between successive durations to draw comparisons between rhythmic structures in music. Equation (4) shows the formula for nPVI, with d^k representing the duration of the interval k. m represented the number of intervals. The use of the formula standardised the difference between successive intervals in relation to their average length [29].
- **Pulsation Clarity** – The strength of the rhythmic pulse in a piece of music

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{3}$$

$$\text{nPVI} = \frac{100}{m-1} \times \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right| \tag{4}$$

A quantitative survey was also conducted and contained a total of 10 questions all containing the question *"Rate the likelihood that this piece was composed by Johannes Brahms as opposed to being generated by AI"* In random order, 5 pieces contained the works of Brahms and 5 were generated by the models. Answers consisted of a Likert Scale of 5 values ranging from *Definitely generated by AI* to *Definitely generated by Brahms.* Members of the Irish Defence Forces School of Music, RTÉ Concert Orchestra and National Symphony Orchestra were recruited for the survey to provide professional expertise in the subject. Participants were selected based on their extensive experience in classical music, including familiarity with Brahms' works, having performed his pieces in the past.

Using IBM's *SPSS* software, quantitative values from the MIR functions *MIRToolbox* and *MIDIToolbox* were tested to obtain p-values. Various Independent-Samples T-Tests and Hotelling's T^2 tests were implemented to evaluate musical variables to determine if there was statistical indistinguishability between Brahms' piano works and the generated pieces from the AI models.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

D. Ethical Considerations

Several ethical considerations were adhered to in order to correctly conduct research including copyright issues and collection of data from survey participants A total of 67 pieces of music were obtained for offline manipulation for the MIDI dataset. According to German Federal Law Gazette, copyright protection for musical and artistic works expired 70 years after the death of the creator. With Brahms passing away in 1897, his compositions therefore resided in the public domain. The use of MIDI files also prevented any issues with performers rights as recordings of Brahms' works were not being used. All the neural network models obtained for testing were open-source and ran under the Apache 2.0 License. MIR tasks were performed using the *MATLAB* functions *MIRToolbox* and *MIDIToolbox*. To utilise the tools, *MATLAB* had to be downloaded free of charge under the GNU General Public License. Ethical considerations were vital when collecting data from personnel for the quantitative survey. Participation for the survey was voluntary and those who chose to partake were informed of the purpose of the study. No personal information was required from participants and the confidentiality of participants was guaranteed from the designer of the survey. The results of the survey were not tampered with and therefore were accurate.

III. EVALUATION

Overfitting issues were observed during training, with the LSTM and Perceiver AR models initially replicating the training material excessively instead of generating unique motifs. To mitigate this, the dataset was further augmented by transposing musical phrases and altering rhythms through quantisation to introduce additional variations in tempo and phrasing. Experimenting with different temperature values for each model allowed for control over the balance between simplicity and randomisation. With lower temperatures producing more basic outputs that resembled the training data, while higher temperatures encouraged greater diversity but could lead to compositions with less structure. Table 1 shows the model performance metrics stating that the transformer model with relative self-attention scored the best training loss and accuracy with scores of 0.015 and 0.995 respectively. The recurrent neural network scored worst in terms of training accuracy with a score of 0.628.

Through quantitative experimentation and a survey, the neural network models utilised for the project were evaluated extensively in a numerical form and through professional human judgement in order to confirm or refute the research hypothesis that neural network models could generate pieces of music with statistically indistinguishable musical characteristics and emotion as Brahms. To conduct a fair experiment, each model had to generate a two-minute-long piece which contained 300 prime tokens (30 seconds) from the beginning of six of Brahms' piano works. These generated pieces were then compared with the first two

TABLE 1. TRAINING LOSS AND ACCURACY SCORES

Model	Training Loss	Training Accuracy
Recurrent Neural Network	0.550	0.628
RNN with Long Short-Term Memory	0.154	0.983
Transformer w/ Self-Attention	0.737	0.939
Transformer w/ Relative Attention	0.447	0.985
Transformer w/ Relative Self-Attention	**0.015**	**0.995**
Transformer w/ Local Windowed Attention	0.015	0.993
Sparse Transformer	0.612	0.803
Perceiver AR	0.825	0.774

minutes of the original Brahms piece to evaluate the evolution of the generated pieces and determine their ability to maintain the style and structure of Brahms. MIR evaluation concluded that the Transformer models with self-attention, local windowed attention and relative global attention performed best in generating music most similar to the Brahms original.

Several statistical tests were conducted on the best performing models to obtain p-values to test the research hypothesis. The variables Entropy, nPVI, Global Energy, Mean Roughness and Pulsation Clarity were tested with an Independent-Samples t-test and the variables Duration Distribution and Pitch Class Distribution being tested with a Hotelling's T^2 test. Based on guidance from Leard Statistics, the choice of test for each variable depended on whether the variable required a joint test. The independent-samples t-test was not suitable in those cases, whereas Hotelling's T^2 test was. Testing concluded that no statistically significant difference was found with most of the variables therefore supporting the alternative hypothesis that neural network models can produce music that is indistinguishable from Brahms. However, the variables Entropy and Global Energy were deemed to contain statistical significance within them stating that further work must be done to improve complexity, uncertainty and energy to a similar level to Brahms. Using *MIRToolbox*, the waveforms of the generated pieces were evaluated.

The entropy of distributions was evaluated to measure uncertainty or unpredictability in succeeding phrases. Table 2 shows how to models performed in terms of entropy from using 300 prime tokens of Brahms' *2 Rhapsodies No.1* to determine which model scored most like the original piece. Results showed that the Transformer model with Local Windowed Attention scored most similar with a score of 0.9214 compared to the original piece's 0.9115.

The transformer model with Local Windowed Attention was then further tested against the other Brahms original tracks to obtain more Entropy values. Table 3 shows the comparison of values with each piece.

With all this data, an independent-samples t-test was performed to determine if there was statistical significance between the Brahms and Transformer model's generated music. First, a Levene's Test for Equality of Variances was conducted to determine if equal variances should be

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE 2. ENTROPY VALUES BETWEEN BRAHMS AND AI MODELS

Model	Entropy	Difference
Brahms 2 Rhapsodies No. 1 – Original	***0.9115***	/
RNN	0.8802	0.0313
LSTM	0.8899	0.0216
Transformer w/ Self-Attention	0.9247	0.0132
Transformer w/ Relative Attention	0.9007	0.0108
Transformer w/ Local Windowed Attention	**0.9214**	**0.0099**
Transformer w/ Relative Global Attention	0.9241	0.0126
Sparse Transformer	0.9310	0.0195
Perceiver Transformer	0.9241	0.0126

TABLE 3. ENTROPY VALUES BETWEEN BRAHMS AND TRANSFORMER MODEL WITH LOCAL WINDOWED ATTENTION

Piece of Music	Type of output	Entropy
2 Rhapsodies No. 1	***Brahms***	***0.9115***
	Transformer w/ Local Windowed Attention	0.9214
2 Rhapsodies No. 2	***Brahms***	***0.9140***
	Transformer w/ Local Windowed Attention	0.9214
3 Intermezzi No.1	***Brahms***	***0.9135***
	Transformer w/ Local Windowed Attention	0.9168
4 Ballads No. 4	***Brahms***	***0.9094***
	Transformer w/ Local Windowed Attention	0.9167
7 Fantasias No. 7	***Brahms***	***0.9135***
	Transformer w/ Local Windowed Attention	0.9192
Hungarian Dances No. 2	***Brahms***	***0.9163***
	Transformer w/ Local Windowed Attention	0.9133

assumed for the t-test. With a scored p-value of 0.401 being much greater than 0.05, equal variances were assumed. Since both groups containing Brahms' original works and generated pieces from the transformer models were being evaluated, a two-sided p-value was observed. The independent-samples t-test stated a value of 0.012 indicating that there was statistical significance between the Brahms and AI generated music groups. The audio waveforms for the Brahms and AI generated pieces were analysed using *MIRToolbox* to discover the reason for the statistical significance. An onset curve depicted a greater variance in the detection of successive notes in the AI generated pieces therefore giving a higher entropy value.

Figure 3 shows the brightness curve of Brahms' *2 Rhapsodies No. 1* alongside the generated piece from the transformer model with local windowed attention. The generated piece again displays a greater variance of frequencies, resulting in a higher entropy score. The entropy results show that the generated pieces provided too much variation and complexity in succeeding phrases when compared to the Brahms originals.

Statistical significance was also recorded in Global Energy between the Brahms pieces and the best scoring model in this category, the Transformer model with Self-Attention. In this case, the Levene's Test for Equality of Variances calculated a p-value of 0.0002 meaning that equal variances should not be assumed. Therefore, the two-sided p-value of 0.035 was provided, indicating there was statistical significance between Brahms and AI.

The difference in global energy is depicted in Figure 4 where the temporal evolution curve reveals a much greater variance in the generated piece compared to Brahms'

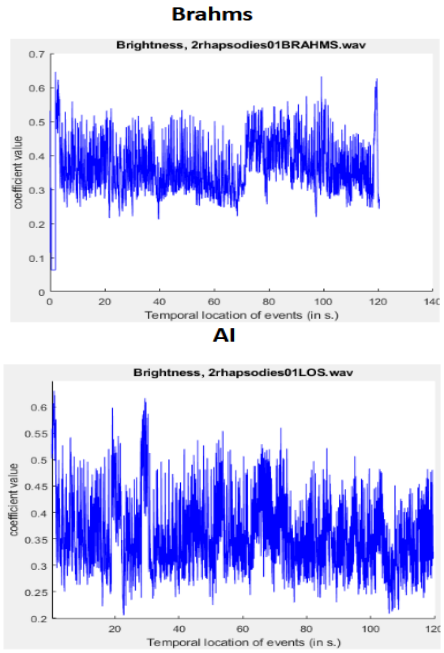


Figure 3. Brightness Curve between Brahms and AI

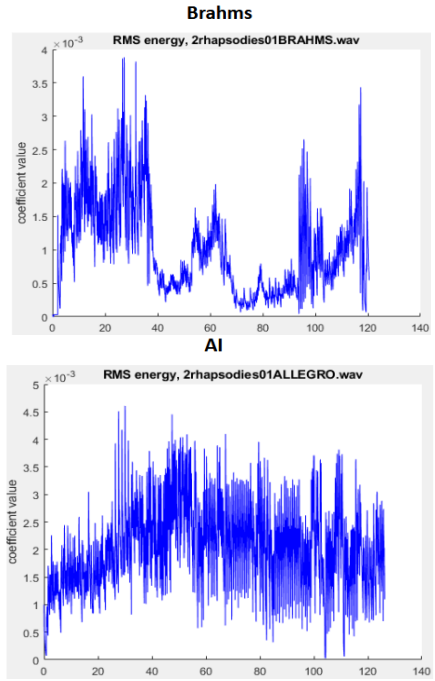


Figure 4. Global Energy Between Brahms and AI

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

original work. While Brahms' piece maintains a steady flow of increase and decrease of tension, the generated piece has much greater variations in timbre and harmonics throughout. The evaluation of these waveforms determined that the entropy and global energy values may have been much higher than Brahms' works due to the *MuseNet* inspired workflow the transformer models undertook causing the pieces to be generated in blocks and therefore sound unnatural.

Although there was statistical significance recognised for global energy, the independent-samples t-test did not recognise any difference in sensory dissonance. With Brahms being considered a more conservative composer who involved little dissonance, or clashing sounds, in his music, it was important to replicate this. Figure 5 shows that the generated piece composed the same methods as Brahms by answering to dissonance with a harmonic resolvement. This can be seen with the sudden peaks in the coefficient value being resolved to lower roughness values.

Figure 6 displays the Pitch Class Distribution in the form of a box plot for Brahms' *3 Intermezzi No.1* and the generated piece from the transformer model with local windowed attention. With the piece being in the key of D# Major, which has an enharmonic equivalent of C minor, in order to stay within the key signature, the majority of the notes must be from the following triads:

- D# Major – D#, G, A#
- C Minor – C, D#, G

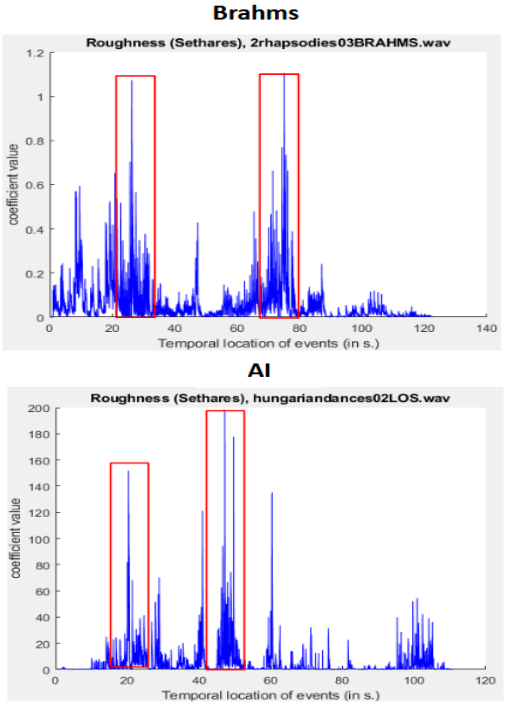


Figure 5. Sensory Dissonance between Brahms and AI

The pitch class distribution showed that both the original Brahms piece and AI generated kept within these guidelines, with particular emphasis being placed on the notes D# and G, as they feature in the triads of both D# major and C minor. Although the transformer model focused on the notes within the two triads to ensure consistency in key signature, it was apparent that the model was reluctant to incorporate accidentals to further add colour to the piece. The use of extended chords was a key factor of the romantic period in which Brahms lived in and it was an era that bridged the gap between classical and modern music. *MIRToolbox* was able to identify both the Brahms and AI pieces to be in the key of D# major. And with the function *mirmode,* it identified that the generated piece scored a higher probability of being in a major key that the Brahms original.

The chromagram at Figure 7 visualises the pitch distribution throughout the generated piece, showing that the D# major triad was used frequently throughout, meaning that the AI piece kept firmly within the major key. Although it was a positive that the generated piece was able to keep within the key signature for longer generated sequences, it does show an inability to evolve melodically into different harmonics.

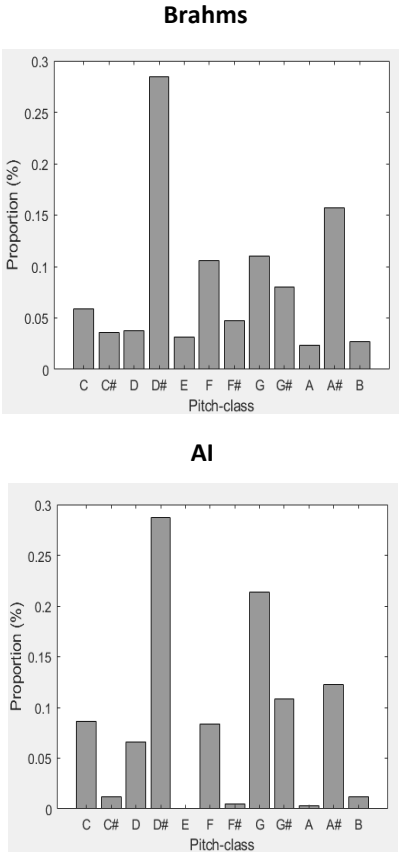


Figure 6. Pitch Class Distribution between Brahms and AI

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

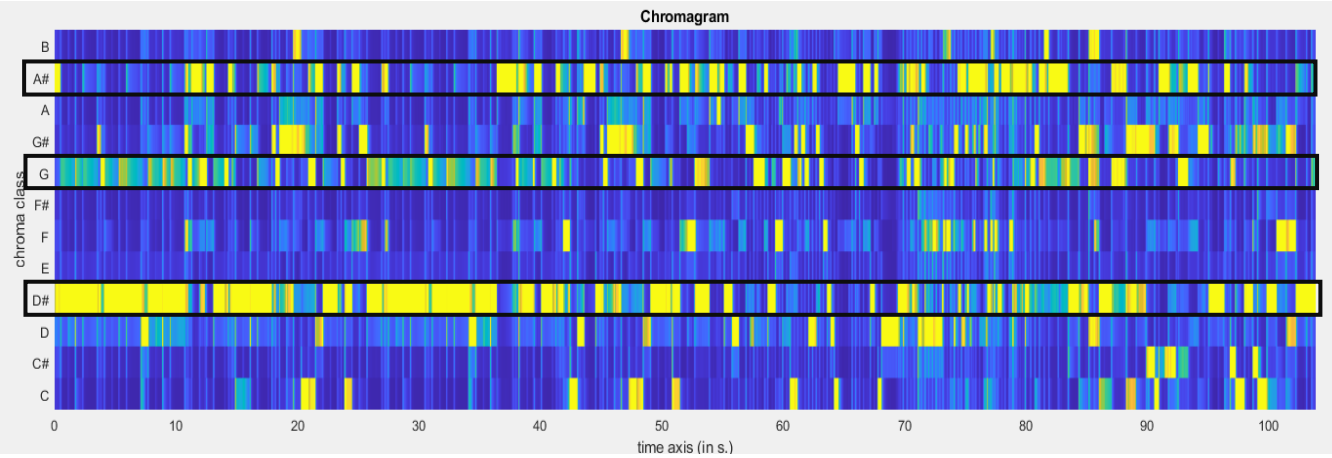


Figure 7. Chromagram showing the pitch class frequency throughout the generated piece by transformer model with local windowed attention

Figure 8 shows the Pulsation Clarity between Brahms and AI. The independent-samples t-test did not recognise any statistical difference between the two therefore the generated music from the transformer model with relative self-attention produced indistinguishable pulse clarity characteristics from Brahms. Pulse clarity did not necessarily mean that the piece must be perfectly in tempo. With the Romantic period being an expressive music era, compositions often involved many tempo pushes and pulls, also known as *rubato.* Therefore, it was important for the AI model to generate the same pulsation characteristics as Brahms.

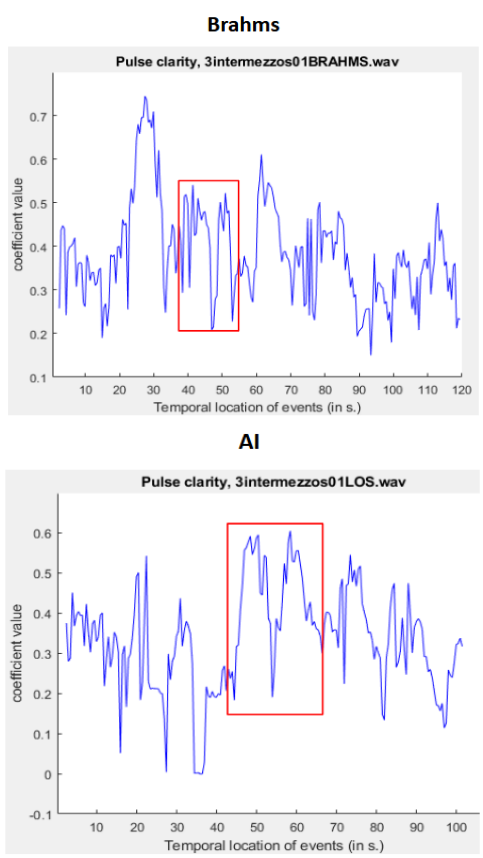


Figure 8. Pulse Clarity between Brahms and AI

The figure shows how the transformer model generated similar pulsation and rhythmic characteristics to the original Brahms pieces with the higher values indicating a strong pulse followed by dips in coefficient value before regaining its original value. This could have been a sign of the music taking rubato into effect.

A quantitative survey was also conducted to obtain human evaluation on 30 second clips of generated pieces from the models against the original works of Brahms. A total of 56 participants featuring only professional musicians, composers and conductors, displayed difficulty in recognising distinction between the Brahms and generated pieces provided, with the majority incorrectly identifying two of the generated pieces as one of Brahms' own works.

Table 4 shows the percentages for each of the questions alongside whether the track was that of Brahms or AI. The total score was also calculated with the number of responses per answer being multiplied dependant on how similar to Brahms it was scored with *Definitely generated by AI* scoring 1 and *Definitely generated by Brahms* scoring 5.

TABLE 4. QUANTITATIVE SURVEY RESULTS

	AI	Probably AI	Unsure	Probably Brahms	Brahms	Total Score
Brahms	3.64%	5.45%	3.64%	36.36%	**50.91%**	**237**
Brahms	12.50%	25.00%	3.57%	**37.50%**	21.43%	**185**
Brahms	0%	23.21%	7.14%	**39.29%**	30.36%	**211**
Brahms	17.86%	**44.64%**	3.57%	28.57%	5.36%	**145**
Brahms	7.27%	7.27%	18.18%	**41.82%**	25.45%	**207**
AI	23.21%	**32.14%**	12.50%	23.21%	8.93%	**147**
AI	10.71%	19.64%	19.64%	**28.57%**	21.43%	**185**
AI	5.36%	**41.07%**	14.29%	32.14%	7.14%	**165**
AI	14.55%	**34.55%**	5.45%	32.73%	12.73%	**165**
AI	7.27%	30.91%	7.27%	**36.36%**	18.18%	**183**

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

This scoring design was intentional so that uncertainty was treated as a reward for the AI models as they still had not been identified as not Brahms by professionals.

An Independent-Samples T-Test was conducted which stated that there was no statistically significant difference between the Brahms and generated pieces therefore supporting the alternative hypothesis by that neural network models trained with an augmented and pre-processed dataset could generate music at the level of musicality and emotion as of Brahms so much that through a quantitative survey the difference could not be identified by professional musicians, composers and conductors.

IV. Conclusion and Future Work

The work conducted in this paper differed to previous studies as it trained various neural network models with a dataset containing the piano works of Brahms, an important figure of the Romantic Period in Classical Music. By doing this, a gap in the research was addressed by analysing a very important composer in an era of classical music were harmonic and rhythmic structures began to diverge from the traditional aspects of Renaissance and Classical Period music while also bridging the gap between traditionalism and modernism. While the literature review stated a gap in the research that previous models struggled with complex motifs and harmonic cadences of romantic period piano music. This paper found that statistical testing in various musical categories stated there was not statistical significance between the Brahms and AI generated pieces. A quantitative survey containing participants who were educated in the subject mistook two of the neural network models generated pieces as Brahms' own works suggesting the model's ability to generate pieces of music with the complexity in rhythmic and harmonic characteristics of Brahms.

The results from the research carried out suggest that transformer models with self-attention, relative self-attention and local windowed attention were able to generate various characteristics to a statistically indistinguishable level from Brahms with a dataset of his piano works by utilising an augmentation pipeline and various preprocessing techniques. A couple of musical characteristics however proved to be statistically significant to Brahms, these being entropy and global energy. This concludes that while the transformer models are able to replicate a vast amount of Brahms' compositional traits, it still falls behind in reproducing the rhythmical and harmonical complexities, uncertainties and global energy levels of Brahms' works. The possibility of increasing the dataset to the orchestral, ensemble and choral works of Brahms could greatly increase the abilities of generated music from just solo piano works. This also could adhere to limitations regarding a small dataset and therefore improve accuracies in entropy and global energy from the generated pieces.

While participants noted difficulty in distinguishing between the Brahms and AI pieces, some commented that the use of MIDI files made all the music sound robotic and therefore made it even more challenging to differentiate between the two. Future work could focus on converting the generated pieces into musical notation and have a professional pianist perform them. This would enable an experiment to assess the generated music on an acoustic piano, the instrument it was originally intended for. The present study included generated music that used a short input sequence derived from the Brahms dataset. Future work will focus exclusively on cold-start generation to ensure all evaluated pieces represent novel compositional output.

References

[1] J. Doherty and B. Tierney, "The Generation of Piano Music in the Style of Johannes Brahms Using Neural Network Architectures", *In Proc of the International Conference on Creative Content Technologies,* no. 17, pp. 7-12, IARIA, 2025.

[2] J. D. Fernández and F. Vico, "AI Methods in Algorithmic Composition: A Comprehensive Survey", *Journal of Artificial Intelligence Research*, no. 48, pp. 513-582, 2013. https://doi.org/10.48550/arXiv.1402.0585

[3] K. Zheng, R. Meng, C. Zheng, X. Li, J. Sang, J. Cai. and J. Wang, "EmotionBox: a music-element-driven emotional music generation system using Recurrent Neural Network", *Frontiers in Psychology*, no. 13, pp. 1-14, 2021. https://doi.org/10.3389/fpsyg.2022.841926

[4] R. Child, S. Gray, A. Radford and I. Sutskever, "Generating Long Sequences with Sparse Transformers", *PsyArXiv,* no. 1, pp. 1-10, 2019. https://doi.org/10.48550/arXiv.1904.10509

[5] C. Hawthorne, A. Jaegle, C. Cangea, S. Borgeaud, C. Nash, M. Malinowski, S. Dieleman, O. Vinyals, M. Botvinick, I. Simon, H. Sheahan, N. Zeghidour, J. B. Alayrac, J. Carreira, and J. Engel, "General-purpose, long-context autoregressive modelling with Perceiver AR", *In Proc of the International Conference on Machine Learning*, no. 39, pp. 1-24, 2022. https://doi.org/10.48550/arXiv.2202.07765

[6] E. Deruty, M. Grachten, S. Lattner, J. Nistal and C. Aouameur, "On the Development and Practice of AI technology for Contemporary Popular Music Production", *Transactions of the International Society for Music Information Retrieval,* no. 5, pp. 35-49, 2022. https://doi.org/10.5334/tismir.100

[7] O. Lartillot, P. Toiviainen, P. Saari and T. Eerola, "MIRtoolbox", 2007. Accessed: Mar. 15, 2024. [Online]. Available: https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox

[8] T. Eerola and P. Toiviainen, "MIDI Toolbox: MATLAB Tools for Music Research", 2004. Accessed: Apr. 22, 2024. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6e06906ca1ba0bf0ac8f2cb1a929f3be95eeadfa

[9] S. Oore, I. Simon, S. Dieleman, D. Eck and K. Simonyan, "This Time With Feeling: Learning Expressive Musical Performance", *Neural Computing and Applications,* no. 32, pp. 995-967, 2018. https://doi.org/10.1007/s00521-018-3758-9

[10] J. A. Franklin, "Jazz Melody Generation from Recurrent Network Learning of Several Human Melodies", *In Proc of the International Florida Artificial Intelligence Research*

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Society Conference, no. 18, pp. 57-62, 2005. https://cdn.aaai.org/FLAIRS/2005/Flairs05-010.pdf

[11] FluidSynth, "pyFluidSynth", 2014. Accessed: Apr. 14, 2024. [Online]. Available: https://github.com/nwhitehead/pyfluidsynth

[12] TensorFlow, "Generate music with an RNN", n.d. [Online]. Available: https://www.tensorflow.org/tutorials/audio/music_generation

[13] C. Leja, "Making Music With Machine Learning: Introduction to Magenta Music.js," 2020. Accessed May 27, 2024. [Online]. Available: https://christopher-leja.medium.com/making-music-with-machine-learning-introduction-to-magenta-music-js-30b4c9dc6279

[14] IBM, "What are recurrent neural networks?", n.d. Accessed: Mar. 5, 2024 [Online]. Available: https://ibm.com/topics/recurrent-neural-networks

[15] J. Frost, "Mean Squared Error (MSE)," n.d. Accessed: May 11, 2024. [Online]. Available: https://statisticsbyjim.com/regression/mean-squared-error-mse/

[16] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation,* no. 9, pp. 1735-1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735J.

[17] C. Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu and D. Eck, "Music Transformer: Generating Music with Long-Term Structure", *In Proc of the International Conference on Learning Representations,* no. 17, pp. 1-14, 2018. https://doi.org/10.48550/arXiv.1809.04281

[18] A. Sigalov, "Project Los Angeles", 2019. Accessed: Apr. 3, 2024. [Online]. Available: https://github.com/asigalov61

[19] A. K. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All you Need", *Advances in Neural Information Processing Systems,* no. 31, pp. 1-15, 2017. https://doi.org/10.48550/arXiv.1706.03762

[20] S. Shankar, "Understanding Google's *"Attention is all you Need"* Paper and its Groundbreaking Impact", n.d. Accessed: Mar. 22, 2024. [Online]. Available: https://alok-shankar.medium.com/understanding-googles-attention-is-all-you-need-paper-and-its-groundbreaking-impact-c5237043540a

[21] A. K. Huang, P.A. Szerlip, M. E. Norton, T.A. Brindle, Z. Merritt and K.O. Stanley, "Visualizing music self-attention", *In Proc of the Conference on Neural Information Processing Systems,* no. 32, pp. 1-5, 2018. https://openreview.net/pdf?id=ryfxVNEajm

[22] P. Shaw, J. Uszkoreit and A. Vaswani, "Self-Attention with Relative Positional Representations," *In Proc of the 2018 Conference of the North American Chaper of the Association for Computational Linguistics: Human Language Technologies,* no. 2, pp. 464-468, 2018. https://doi.org/10.18653/v1/N18-2074

[23] Z. Liu, Y. Lin, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swim Transformer: Hierarchical Vision Transformer using Shifted Windows," *In Proc of the IEEE/CVF International Conference on Computer Vision,* no.23, pp. 9992-10002, 2021. https://doi.org/10.48550/arXiv.2103.14030

[24] OpenAI, "Generative modelling with sparse transformers," 2019. Accessed: May 16, 2024. [Online]. Available: https://openai.com/index/sparse-transformer

[25] asigalov61. "TMIDIX.py," n.d. Accessed: May 18, 2024. [Online]. Available: https://github.com/asigalov61/Giant-Music-Transformer/blob/main/TMIDIX.py

[26] S. Y. Huang and Y. H. Yang, "Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions," *In Proc of the ACM International Conference on Multimedia,* no. 28, pp. 1180-1188, 2020. https://doi.org/10.1145/3394171.3413671

[27] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals and J. Carreira, "Perceiver: General Perception with Iterative Attention," *In Proc of the International Conference on Machine Learning,* no. 38, pp. 1-43, 2021. https://doi.org/10.48550/arXiv.2103.03206

[28] M. Rayden, "What is RMS in audio world?," 2024. Accessed June 2, 2024. [Online]. Available: https://majormixing.com/what-is-rms-in-audio-world/

[29] A. D. Patel and J. R. Daniele, "An empirical eomparison of rhythm in language and music," *Cogntion,* no. 87, pp. 35-45, 2002. https://doi.org/10.1016/S0010-0277(02)00187-7

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Extending the Composable Architecture Framework by Means of Normalized Systems Theory

Geert Haerens
Antwerp Management School, Belgium
Engie nv, Belgium
Email: geert.haerens@engie.com

Herwig Mannaert
University of Antwerp, Belgium
Email: herwig.mannaert@uantwerpen.be

***Abstract*—In a fast-evolving world, companies require IT solutions that allow them to adapt swiftly to changing conditions. In 2020, Gartner introduced the Composable Architecture Framework as a guiding principle for creating application landscapes that are easily composable and recomposable, thereby supporting change. The Normalized Systems theory is about the creation of evolvable modular software. The concepts presented in the Composable Architecture Framework resonate with Normalized Systems. A closer analysis of the framework through Normalized Systems theory reveals that Gartner's framework lacks precision. As such, following the guidance of the framework will insufficiently protect companies from change, both outside and inside the organization. By extending the Composable Architecture Framework with Normalized Systems theory, a more evolvable framework emerges.**

***Keywords*—Normalized Systems Theory; Composable Architecture; Packaged Business Capabilities.**

I. Introduction

This paper is an extended version of [1]. We elaborate on the concepts behind the Composable Architecture Framework (CAF) and Normalized Systems (NS) theory and extend our criticism of the original paper, providing guidance on how to improve CAF.

For many years now, a death wish toward the monolithic application has been declared. Monolithic applications are difficult to change and unsuitable in our fast-moving world. Many paradigms have been proposed over the years to split applications into smaller, modular parts. With the rise of the Internet and faster network speeds, the physical distribution of those parts has become a reality. The Distributed Computing Environment (DCE) [2] proposed modules encapsulating functionality that could be activated via Remote Procedure Calls (RPC). An essential benefit to splitting applications into smaller, independent callable modules is re-use. This resonates with McIlroy's dream, expressed during the 1968 NATO conference on Software Engineering, where *" ... I expect families of routines to be constructed on rational principles so that families fit together as building blocks. In short, [the user] should be able to regard components as black boxes safely."*.

SAP ERP (Enterprise Resource Planning), for instance, uses this approach to allow calling SAP functions from other systems via Remote Function Calls (RFCs). SAP re-baptized RFC to BAPI, Business Application Programming Interface, to focus even more on encapsulated functionality. In the past couple of years, we have seen a shift from encapsulated functionality toward technology, meaning that today, the talk of the town is about REST APIs, message queues, and event systems as a means of implementing integration between modules but without paying attention to the functionality.

In 2020, Gartner [3] introduced the notion of PBC to create Composable Applications. They proposed the CAF [3] as a guide for properly encapsulating functionality and using technologies that enable recomposition and evolution. Their approach connects the previous focus on functionality with today's emphasis on technology.

Normalized System theory (NS) [4], originating in software development, put forward the necessary conditions for the evolvability of modular structures.

When analyzing the CAF with NS, one starts noticing that CAF lacks precision. Following the guidance outlined by CAF will not be sufficient to protect a company's IT systems from change. This brings us to the problem statement we aim to address: How can NS help extend CAF to address the identified weaknesses?

The paper is structured as follows: In Section II, we will introduce Gartner's Reference Architecture for Composable Business Technology, followed by Section III, that introduces NS. In Section IV, we refer to related work, and in Section V, we analyze and criticize the framework through NS. In Section VI we have our findings validated by a focus group and in Section VII we will propose extensions to CAF to improve the overall robustness against change. The paper is wrapped up in Section VIII.

II. Gartner's Reference Architecture for Composable Business Technology

In 2020, the world was struck by COVID. In addition to human loss and suffering, businesses were severely disrupted by this crisis. Gartner noticed that companies with a modular application approach could adjust swiftly to external conditions by quickly and safely assembling, disassembling and reassembling applications as the world required. In November 2020, Gartner published their Reference Architecture for Composable Business Technology [3], that investigates the conditions required for a platform to facilitate the composition and re-composition of applications. The necessary ingredients for such a platform are PBCs as the essential modules, an application composition experience that allows custom assembly

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

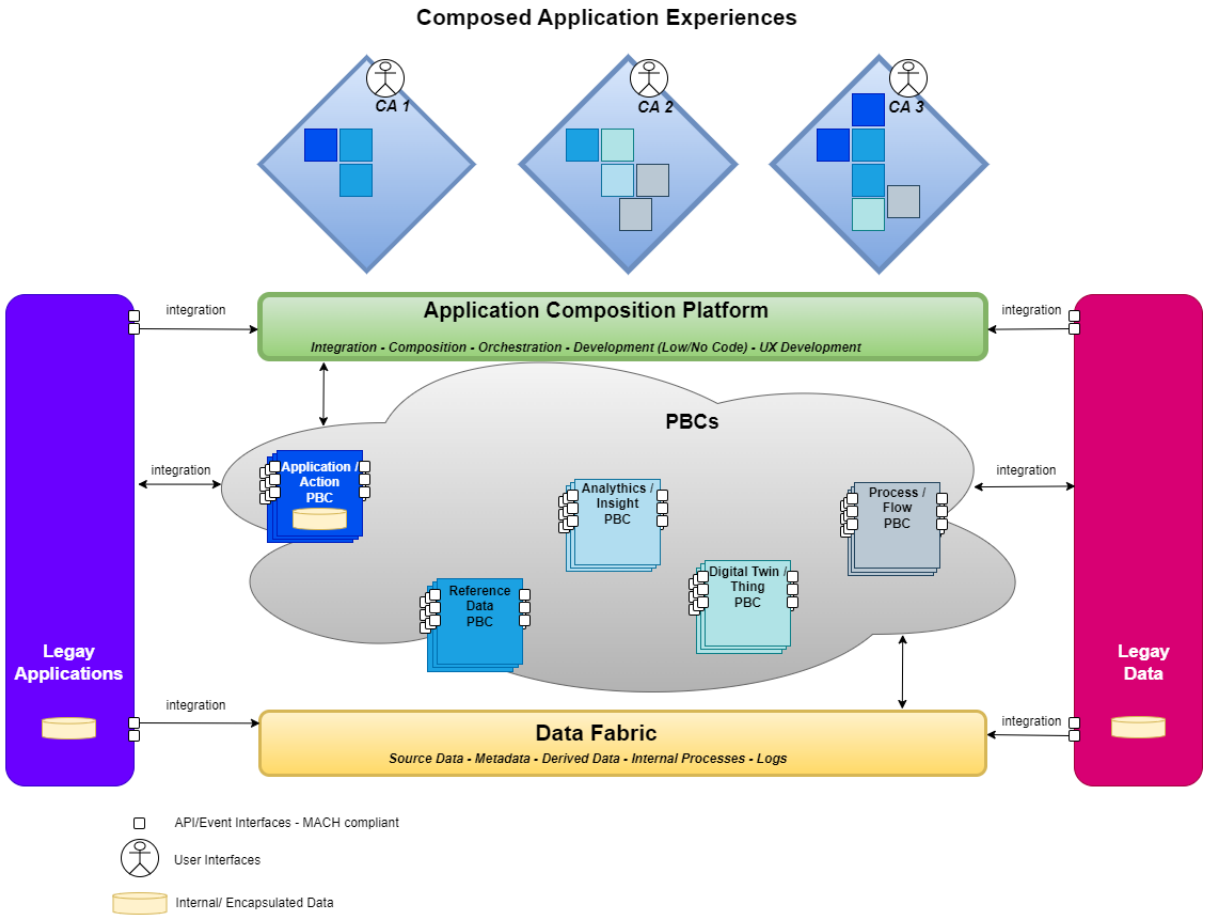


Fig. 1. Overview of the CAF [5].

of PBCs via low-code and no-code, an application composition platform that enables development and deployment of the newly composed applications, and a data fabric that provides easy access to data and analytics.

The PBCs are modules that encapsulate a well-defined business capability (BC) - recognized by the intended business users - and must adhere to the following conditions:

- They are modular and cohesive.
- They are autonomous (can run independently) and have minimal dependencies with external components.
- They allow orchestration as they can realize a process flow across PBCs (via APIs, events, etc.).
- They are discoverable and easily accessible and recognizable by those who require them.

The above definition of a PBC is heavily focused on the ”P” — how it should be packaged — and less on the ”BC” — the business capability. In Section V, we will see that the defining BCs is not straightforward and that the lack of definitions and guidance from Gartner is a weakness of the PBC concept.

Defining the right level of business capability granularity is challenging (too large = monolith, too small = more complex to identify).

Gartner provides further guidance on the PBC definitions by defining PBC types.

There is the Application Type PBC that encapsulates both data and functions related to a well-defined business capability. Application Type PBC can be used to create fully expressed (=autonomous and encapsulating a full context) PBCs or basis business function PBC (= not autonomous and encapsulating a part of a context). Application Type PBC can create pseudo-PBCs that act as APIs to existing monolithic applications. They encapsulate the legacy application, allowing them to participate in PBC composition, but they lack some of the flexibility in true PBCs regarding evolvability.

Reference-type PBCs allow encapsulated access to data in the data fabric. They can access master data, metadata, or any data instance that represents a business object or a physical data container.

The Insight Type PBC allows the encapsulation of data analytics processes. It can perform analytic operations or apply AI models (ML, Deep learning, etc.) to data in the data fabric.

The Thing Type PBC encapsulates physical-world entities. It can be used to access and manipulate data from the real world (IoT).

There is the Flow Type PBC, that combines different PBC

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

in a specific order. Flow Type PBCs facilitate the creation of orchestrated PBCs that encapsulate a process.

The PBCs are activated via event channels and called via APIs. They can also provide different optional user interfaces (web, mobile, etc.).

In Table I, Gartner contrasts the organization of an application landscape in PBCs (combined with a composition and deployment platform) with traditional application landscapes. Figure 1 provides an overview of the CAF.

Gartner analyzed the main change drivers in their papers: adding new business capabilities and their associated PBCs. Gartner further emphasizes that a platform that conforms to the above specifications is insufficient. Business and IT must work closely together to define and implement the required business capabilities, requiring what Gartner calls fusion teams. These teams are essential in creating business-IT alignment and, thus, value creation.

To start with a composable architecture strategy, one must first know what PBCs are already in the company. They may already be present as fine-grained PBCs, or via applications that aggregate PBCs and are accessible via APIs. It is vital that investments in future technologies that can be used for PBCs can be used for that purpose. The individual PBCs must be accessible via APIs, not aggregated. For example, a SaaS solution integrating multiple PBCs should allow the specific use of platform PBCs without depending on the other PBCs.

New applications are composed of assembled PBCs, allowing faster delivery and updating. Updating a PBC in the catalogue should update all applications with that PBC as an active module.

In summary, the Gartner CAF *"presents the reference model for developing business applications that are modular, composable, easily adapted and ready for change."* [3].

III. Fundamentals of NS theory

Software should be able to evolve as business requirements change over time. In NS theory [6], the lack of Combinatorial Effects measures evolvability. When the impact of a change depends not only on the type of the change but also on the size of the system it affects, we talk about a Combinatorial Effect. The NS theory assumes that software grows indefinitely and undergoes unlimited changes over time. Under those conditions Combinatorial Effects harm software evolvability.

NS theory is built on the principles of classical system engineering and statistical entropy. In classic system engineering, a system is stable if it has bounded input and bounded output (BIBO). NS theory applies this idea to software design, as a limited change in functionality should cause a limited change in the software. In classic system engineering, stability is measured at infinity. NS theory considers infinitely large systems that undergo infinitely many changes. A system is stable for NS if it does not have Combinatorial Effects, meaning that the effect of change only depends on the kind of change and not on the system size.

In the context of composable architecture, NS theory provides a lens for evaluating whether PBC-based systems will remain stable as they evolve. As Gartner claims, the CAF should result in more evolvable and thus easier to change systems; compliance with NS seems necessary.

NS theory proposes four theorems and five extendable elements as the basis for creating evolvable software through the pattern expansion of these elements. The theorems are proven formally, giving a set of required conditions to follow strictly to avoid Combinatorial Effects. The NS theorems have been applied in NS elements. These elements offer a set of predefined higher-level structures, patterns, or "building blocks" that provide a clear blueprint for implementing the core functionalities of realistic information systems, following the four theorems.

The NS elements are the building blocks needed for any modular architecture, and thus also for the CAF.

A. NS Theorems

NS theory [6] is based on four theorems that dictate the necessary conditions for software to be free of Combinatorial Effects.

- Separation of Concerns
- Data Version Transparency
- Action Version Transparency
- Separation of States

Violating any of these four theorems will lead to Combinatorial Effects and, thus, non-evolvable software under change.

B. NS Elements

Consistently adhering to the four NS theorems is challenging for developers. First, following the NS theorems leads to a fine-grained software structure, that introduces some development overhead and may slow the development process. Second, the rules must be followed constantly and robotically, as a violation will introduce Combinatorial Effects. Humans are not well-suited for this kind of work. Third, the accidental introduction of Combinatorial Effects results in an exponential increase in rework.

Five expandable elements [7] [8] were proposed, that make the realization of NS applications more feasible. These elements are carefully engineered patterns that comply with the four NS theorems and that can be used as essential building blocks for various applications: data element, action element, workflow element, connector element, and trigger element.

- **Data Element**: the structured composition of software constructs to encapsulate a data construct into an isolated module (including get- and set methods, persistency, exhibiting version transparency, etc.).
- **Action Elements**: the structured composition of software constructs to encapsulate an action construct into an isolated module.
- **Workflow Element**: the structured composition of software constructs describing the sequence action elements should be performed to fulfil a flow into an isolated module.
- **Connector Element**: the structured composition of software constructs into an isolated module, allowing external

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I
CONTRASTING TRADITIONAL WITH PBC APPLICATION LANDSCAPES (FROM [3])

Criteria	Traditional Applications	PBCs
Primary value	Business capability	Business capability
Primary access	User Interface (UI)	Programmatic Interface (API, event)
Scope	Many business objects	One business object
Internal architecture	monolith or modular	monolith or modular
Designed for	Business	Business and IT
Design priority	Stability	Agility
Delivered value	Business solutions	Recomposable business solutions
Production style	Project	Product
Essential tools	Customization included	Composition, added cost
Required IT skills	Customization, low	Composition, high
Cost	Bulk, some "shelfware"	Componentized, tracks value
Governance	Sample	Complex
Internal data	"Owned"	"Owned"
Open for integration/composition	Partially, a secondary priority	Fully, primary design objective

systems to interact with the NS system without calling components statelessly.

- **Trigger Element**: the structured composition of software constructs into an isolated module that controls the system states and checks whether any action element should be triggered accordingly.

The element provides core functionalities (data, actions, etc.) and addresses the Cross-Cutting Concerns (CCCs) that each of these core functionalities requires to function correctly. CCCs cut through every element, requiring careful implementation to avoid introducing Combinatorial Effects.

C. Element Expansion

An application comprises data, action, workflow, connector, and trigger elements that define its requirements. The NS expander is a technology that generates code instances of high-level patterns for a specific application. The expanded code will provide generic functionalities specified in the application definition and will be a fine-grained modular structure that follows the NS theorems (see Figure 2).

The application's business logic is now manually programmed inside the expanded modules at pre-defined locations. The result is an application that implements the required business logic and has a fine-grained, modular structure. As the code's generated structure is NS-compliant, we know it is evolvable for all anticipated change drivers corresponding to the underlying NS elements. The only location where Combinatorial Effects can be introduced is in the customized code.

D. Harvesting and Software Rejuvenation

The expanded code includes predefined areas where changes can be made. To keep these changes from being lost when the application is expanded again, the expander can gather them and re-inject them when re-expanded. Gathering and putting back the changes is called harvesting and injection.

The application can be re-expanded for various reasons. For example, the code templates of the elements can be improved (fix bugs, make faster, etc.), new CCCs (add a new logging

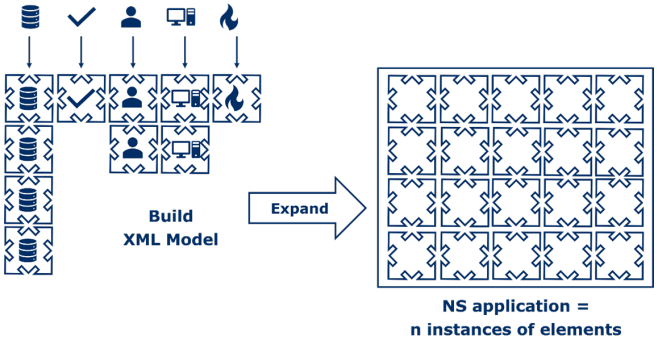


Fig. 2. Requirements expressed in an XML description file, used as input for element expansion.

feature) can be included, or a technology change (use a new persistence framework) can be supported.

Software rejuvenation aims to routinely perform harvesting and injection to ensure that constant enhancements to the element code templates are incorporated into the application.

Code expansion accounts for more than 80% of the application's code. The expanded code can be called boilerplate code, but it is more complex than what is usually meant by that term because it deals with CCCs. Manually producing this code takes a lot of time. Using NS expansion, this time can now be spent on constantly improving the code templates, developing new templates that make the elements compatible with the latest technologies, and meticulously coding the business logic. Changes to the elements can be applied to all expanded applications, giving the concept of code reuse a new meaning. All developers can apply a modification to a code template by one developer across all their applications with minimal impact, thanks to the rejuvenation process (see Figure 3).

IV. RELATED WORK

While searching for relevant literature, we found some papers discussing Composable Architecture. The most relevant publications are, on the one hand, a paper about a possible

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

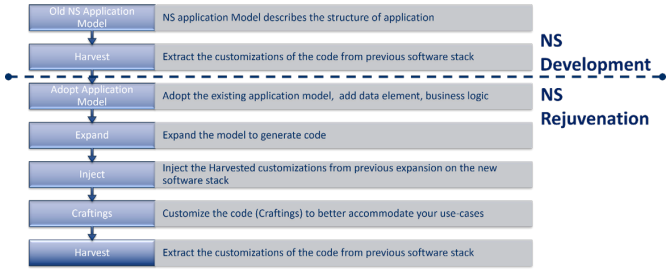


Fig. 3. NS development and rejuvenation.

methodology and demonstration by use case for implementing Composable Architecture, and on the other hand, the book of AW Sheers, ”The Composable Enterprise” [9]. While Scheer [9] tried to rearchitect a complete organization, Ivas [10] provides a methodology to introduce Composable Architecture and demonstrates it with a use case. The paper also provides a good literature review, noting that only a few academic papers discuss Composable architecture (searches on Google Scholar, Web of Science, and Resource Gate). At the same time, thousands of papers are found in the professional literature (via Google), pointing out the need for academic attention to the subject. The paper proposes a methodology consisting of 6 steps (from [10]):

- Understand business drivers and objectives. The first step is to understand the business background of the initiatives, i.e., the rationale, purpose, and scope from the business point of view.
- Understand the holistic scope of the initiative. The second step is to understand what value stream steps and business capabilities from Holistic value delivery are affected by the initiative and how.
- Understand the current situation. Understand and sketch the scope of the current solution (architecture).
- Understand the situation and needs at the enterprise level. Identify if any components can be reused or optimised by this solution at the enterprise level or if there are other future initiatives with the same need.
- Design as API -first Headless PBC (preferably according to MACH [11]). If you need to implement a new service, you should preferably design it according to MACH principles. Otherwise, deliver business change by creating new or optimising existing monolith modules by API-first and Headless MACH principles.
- Implement business-IT aligned PBC solution and consolidate. Implement the agreed business-IT solution and consolidate any old solutions into the new solution that implements the same functionality (business capability).

For an Enterprise Architect, those steps are logical and provide excellent guidance. We argue that, next to the need for academic attention to applying Composable Architecture for (re)introducing functional re-uses, there is a need as well for academic attention to properly operationalizing/implementing Composable Architecture to make the dream of McIllroy a reality and not a nightmare.

For this extended paper, we sought additional work on the CAF. We knew that since the publication of the original paper, there were some conferences, mainly non-academic, that had the topic of CAF on the agenda. We can start by stating that, to the best of our knowledge, we have not found any additional academic research on CAF. We have found a lot of gray literature on the subject, indicating that the concept is gaining traction with practitioners and is becoming part of the heuristic body of knowledge of Enterprise Architecture.

For starters, we looked again at Gartner. In 2025 Gartner starts to associate the topic of composability with Digital Experience Platform (DXP), that focuses on the constant need to adjust to new user journeys to keep and attract customers. The link to the need to constantly adapt, change, and evolve is all variations on the same theme: you need a CAF. In the 2025 Magic Quadrant for DXP [12], Gartner states that by 2026, 70% of organizations will be required to adopt composable DXP technology (as opposed to monolithic suites) and explicitly defines that vendors must offer “modular, API-first approaches, with a set of discrete, task-oriented and independently deployable PBCs.

Secondly, we refer to the Mach Alliance, a consortium advocating Microservices, API-first, Cloud-native, Headless (MACH) architectures, heavily featured composability in 2025. They hosted “The Composable Conference 2025” (April 2025, Chicago) with hundreds of tech leaders, focusing on composable approaches and real-world MACH case studies. The alliance’s insight articles note that composability has “come of age” – e.g., pointing out that “in 2025, composability has become a mandatory requirement” for DXPs and highlighting that many new DXP entrants adopt the PBC approach. The MACH Alliance also offers whitepapers on evaluating and integrating composable solutions, indicating an increasing level of practical guidance available in the industry [13].

Thirdly, we refer to SAP. In early 2025, SAP introduced the concept of Composable ERP, reframing its traditionally monolithic ERP suites into modular components. At the ESICONF 2025 conference [14], SAP’s approach was detailed: they identify PBCs within ERP (e.g., Finance, Inventory, and Logistics as separate capabilities) and allow customers to implement them independently and integrate them as needed. SAP even defined composability similarly to Gartner, as “the ability to select, assemble, and rearrange components to adapt to specific and evolving requirements”. What’s significant is SAP’s explicit breakdown of a PBC’s internal architecture into three layers – User Interface, Application/Process Logic, and Unified Domain Data Model. This provides a more formal metamodel for PBCs in an ERP context, essentially an extension of CAF: it suggests that each PBC should include its own UI and business logic, and should share a standard data model where appropriate. SAP’s messaging around Composable ERP positions it as the next evolution beyond post-modern ERP, aiming for flexibility and the ability to mix and match modules. This can be seen as an “alternative framework” in that it applies composability within a specific

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

enterprise context (ERP) rather than the general application composition of Gartner's CAF. But it's clearly inspired by the same principles and brings them to life in a real product strategy. For practitioners, the SAP case is a strong validation that composable architecture is not just for greenfield digital projects – it's influencing core enterprise systems now. (Other major software providers like Salesforce have similarly started promoting modular, "API-based" architectures in their ecosystems, effectively encouraging composable thinking within their product lines.)

We finish by mentioning that most of the ambiguities that will be discussed in the next section, also pop up in the gray literature (examples: [15], [16], [17], [18], [19]) and that there are ample posts to be found on the web that discuss the issues of defining PBCs, finding the right granularity and using the proper technologies. These topics will be addressed in Section V and will be shortly revisited in Section VI as well.

V. Critical Analyzis of Gartner's Reference Architecture for Composable Business Technology

In this section, we will take the components of the CAF (see Figure 1) and critically analyze them. We start by looking closely at business capabilities and how they should help align business and IT. We continue by discussing the modularity of PBCs and the different types defined by the CAF. We end this section by examining PBCs and CCCs and by delving into the requirements for a PBC platform.

A. Defining Business Capabilities

The concept of business capabilities originates in the resource-based view of companies [20], where it is considered vital to identify resources and capabilities that provide a competitive advantage [21]. This evolved into the idea that companies need to know "what" kind of activities they are undertaking, and gave rise to using the term business capabilities. Although decades have passed since the first mention of capabilities, there is no agreed-upon definition of business capabilities, nor is there consensus on how they should be defined, named, and used.

Table II gives an overview of 4 existing definitions of business capabilities. The first three are from industry sources, while the last is an academic source. There is some overlap, but each has specificities, as outlined in the key differentiators in the table. It is beyond the scope of this paper to provide an overarching definition of a business capability. We refer to [21] for a comprehensive literature overview on the subject. However, it does point to a weakness in the CAF — how to define something that lacks consensus?

Another fundamental issue with business capabilities is that they focus on the "what" rather than the "how". Formulated otherwise, they provide a black-box or a functional view. Traditional/heuristic functional design pays little attention to the fact that moving from function to construction (and vice versa) is not a straightforward, one-to-one mapping. It is an m-to-n mapping and requires a creative process that is not easily caught in an algorithm and requires architectural guidance. This has implications when a company wants to create its business capabilities map, starting from its current structure (how the company is organized). As there is no one-to-one mapping between functions and constructions (and vice versa), one ends up with multiple functions linked to a single construction, and vice versa. To make matters even more complicated, the TAO theory [22] argues that functionality is relative. An object, that is part of the construction, does not have functionality as a property. It only acquires functionality when a subject intends to use an object for a particular purpose. Functionality is thus a relation between subject and object. This implies that when different subjects (persons) build a functional architecture (using business capabilities), they are unlikely to arrive at the same result.

The cornerstone of Gartner's CAF is PBCs. However, as argued above, the definition of business capabilities can be problematic. If the definitions are unclear, then the implementation into PBC is suspect. For some industries, there are business capabilities frameworks such as BIAN [23] for the banking industry or NBility for energy transmission and distribution in the Netherlands [24]. There are companies like Leading Practice [25] that offer capability frameworks for some industries at a price. Still, many sectors lack such frameworks (openly or for pay).

As such, a necessary condition for using the CAF, being well-defined business capabilities, is already a tough nut to crack, and scientific guidance on how to do it is lacking. Some heuristic guidance will be provided in Section VII.

B. Business IT Alignment using PBCs

Gartner considered the business capabilities to be well-defined, shared between business and IT, and as an accurate alignment tool between the two. They refer to "Fusion Teams" as the secret formula to create this shared understanding. Fusion teams are teams that include both business and IT people. Close collaboration between the two will lead to a shared understanding and better solutions. This idea is also present in Agile Frameworks such as SAFe [26], where agile teams are considered multi-functional (mix of business and IT) and have people who know the job best close together, resulting in better designs than those imposed by intentional architecture. Or, stated otherwise, the PBCs are to be built bottom-up rather than defined top-down.

Existing heuristics for developing your business capabilities (see Section VII) treat them as part of the realm of strategy. They are to be aligned with the mission, vision, and strategic objectives. This implies that PBCs are to be built top-down. Such an approach will typically result in a set of business capabilities that are sufficiently understood by IT. The alignment between business and IT is at risk.

From NS, we know the importance of having an anthropomorphic design. This means that module names should reflect something that exists in the real world, as this increases understanding of the module's (intended) function. The same holds

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE II
COMPARISON OF BUSINESS CAPABILITY DEFINITIONS

Source	Definition Summary	Focus	Abstraction Level	Key Differentiator
OMG [27]	Ability or capacity a business possesses or exchanges to achieve a purpose or outcome.	Purpose-driven, exchange of capabilities	High-level, conceptual	Emphasizes exchange and specific purpose.
TOGAF [28]	Ability an organization, person, or system possesses; describes what can be done, not how.	Generic ability, independent of implementation	Very high-level, abstract	Stresses agnosticism to process or structure.
BIZBOK [29]	Ability needed to perform to achieve mission; expressed in terms of outcomes, independent of structure/process.	Mission alignment, outcome orientation	High-level, linked to mission	Highlights mission-critical nature and outcome focus.
Academic [21]	Capabilities link strategy (why) and implementation (how); describe what activities must be performed to achieve objectives, independent of processes or structure.	Strategic alignment, activity-based view	High-level, bridging strategy and execution	Connects strategy to execution explicitly.

for business capabilities. If they are insufficiently fine-grained and abstract, they no longer represent business reality. Still, if they are too detailed, the number of business capabilities is considered too large and thus complex.

C. PBCs and Modularity

Gartner's PBCs are modular, where modularity is defined as "partitioned" into a cohesive set of components [3]. Garter further adds that *"The granularity of PBCs, as with all modular systems, is a common design challenge. Modular components that are too large may be easier to manage, but they are harder to change are use in new compositions. Components that are too small may be easier to assemble but harder to isolate, identify, find or change."* [3]. This statement is vague. What are the objective criteria for being too large and too small? Secondly, as neither is considered a good modular design, what are the requirements for a sound module size? Some of the gray literature mentioned in Section IV considers the PBC to be the optimal size, positioned between the monolith and the microservice. We could dedicate a paper to the discussion of microservice size, but the issue is that an objective criterion is missing.

NS theory explicitly defines the optimal module size to facilitate anticipated changes. A module must be split into smaller modules until each complies with the four NS theorems: SoC, AvT, DvT, and SoS. When all the theorems are met, the optimal size is reached. As these are the necessary conditions for system stability under change, violating them will introduce Combinatorial Effects (CE).

Separation of Concern (SoC) teaches that each PBC should encapsulate a separate change driver. PBCs must be defined as fine-grained and very specific. For example, a possible PBC is Asset Monitoring. The asset could be an IT system (servers, databases) or an OT system (Operational Technology as SCADAs, Turbines, etc.). Although both are assets, secure monitoring for both requires different implementations. This would mean that if we used only one PBC for this, we would need two versions: one for IT and one for OT. It suffices to extend this way of thinking and conclude that insufficiently fine-grained PBCs would lead to multiple versions of the PBC and break the anthropomorphic relation between what something is (a PBC) and how something is done. Such an anthropomorphic relation is a required condition for a PBC platform as PBCs must be easily discoverable (see Section II)

Action version Transparency (AvT) enforces interface stability when the implementation changes. The CAF does not explicitly mention interface stability when the PBC implementation changes. However, it does note the need for technologies that result in loosely coupled systems, such as providing interfaces via APIs and event buses, and using technologies promoted by the MACH Alliance [11]. We will return to this point when discussing the PBC platform. However, AvT should not be limited to technical protocols and implementations. PBC definitions should also address semantic issues related to version transparency. In case a business capability introduces a new feature, a default behavior needs to be defined in case the new feature is not specified.

Data version Transparency (DvT) ensures that changes to data attributes do not affect processing functions that do not use those new attributes. The CAF does not mention this concept; it is left to the data fabric. This paper will not discuss the vagueness or evolvability of concepts such as data fabric.

Separation of State (SoS) required all modules to maintain state and make it externally visible. The CAF does not mention the need to state-keep the PBCs. Similar to AvT, MACH technologies are cited for promoting the use of process Choreography over process orchestration for evolvability. Independent of whether this is valid, choreography (and orchestration) requires state-keeping and exposition to trace process issues back to the failure location. Similar to AvT, statekeeping should also be looked at semantically, i.e., what is the actual business state at specific points in the process flows.

D. PBC Types

Garter introduces PBC Types, such as application PBC, data entry PBC, analytics PBC, Digital Twin, and Process. What is missing is the relation between a PBC and a PBC Type. Is a PBC comprised of multiple PBC types, or do they have a one-to-one relationship? The latter would be strange as to do something, a "what" or a capability, you require things to make it happen, "hows", and the PBC Types are pointing more into the direction of a "how" than in the direction of a "what". The PBC Types have similarities with NS elements.

- Application PBC Type (action) versus the Task Element

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

- Data Entity PBC Type (reference) versus Data Element
- Process PBC Type (flow) vs Flow Element

For NS, an Analytics PBC would be an Application PBC, where the action/task is to perform some analytic operation. A similar reasoning applies to the Digital Twin (thing) PBC, where this would combine data and task elements in NS. We notice Garter's difficulty in addressing concepts at their suitable granularity and abstraction level.

E. PBCs and CCCs

NS recognize that CCCs must be treated as any other change driver and require splitting into different modules and proper encapsulation. PBC focus on business, not technology. The technologies used to implement, run, and deploy the PBC are abstracted, and the only type of guidance regarding technology is to use MARCH-compliant technologies. As stated in the previous subsection, we notice an analogy between PBC Types and NS Elements. NS Elements are there to allow proper encapsulation and separation of CCCs. One would expect something similar in PBCs, but this is not the case. This represents a lack of design criteria for the PBC and could result in proper SoC regarding capabilities/functionalities but zero evolvability for the cross-cutting concern and associated technologies. If CCCs aren't modularized, a PBC might be internally entangled with specific technologies or policies, making it non-evolvable when those concerns change. For example, if every PBC implements its own logging or security, a change in security policy would require modifying many PBCs (a combinatorial effect). As new technological implementations change faster and faster, ignoring these change drivers will introduce CEs. Gartner implies leaving this to the PBC Platform.

F. PBC Platform

Much of the PBC implementation magic is left over to the PBC Platform. It must facilitate the discovery of the available PBC, the composition of PBCs and the deployment of PBCs. We already argued that proper discovery requires anthropomorphising the PBC namespace and its associated granularity to avoid PBC versioning without meaning. In the previous subsection, we argued that Gartner seems to push the treatment of the CCCs toward the PBC Platform. Instead of addressing CCCs at the PBC level, they need to be discussed at the PBC Platform level. This would be a possibility were it not for the fact that, as a selection criterion for a PBC platform, it is currently not mentioned in the CAF

Another crucial aspect of the PBC platform is the design, deployment and running of the PBCs. Gartner sees that a change in a PBC Type would yield an update of all composed applications that use such PBC Type. In NS, updates to the templates that make up the elements are handled through expansion and rejuvenation. The new version of the element template triggers a rejuvenation cycle, updating all instances where it is used. The final step is to deploy the application.

Gartner simplifies this concept with an example of Planning PBC used in two compositions [5]. The update to that Planning PBC would trigger updates to both compositions. The question is, what is being updated? Is it a PBC Type that underlies the Planning PBC? Is it the Planning PBC template, or the code that makes up the PBC? Does this update result in one deployment of the Planning PBC shared by two compositions, or are there two deployments of the Planning PBC? In the former case, it would mean that the Planning PBC has zero customisations (even in terms of low and no-code). Otherwise, it cannot be used in two different composite applications without putting extra differentiation logic into the Planning PBC or the composite applications. In the latter case, it would mean that running instances of Planning PBCs are not identical, and the difference must be managed on the Platform, resulting in different PBC versions. We already discussed the issues related to that topic.

The problem is that there is no clear guidance on how the PBC platform should handle this. The PBC Platform must also conform to the four NS principles, or it will not allow the evolvability Gartner portrayed in its CAF.

VI. Validation By Focus Group

The previous section looked at the components of the CAF and tried to point out possible operationalization issues employing NS and business capabilities theories. To avoid our findings being perceived as overly opinionated, we conducted a small experiment with students at the Antwerp Management School who had just been exposed to NS. The idea was to investigate how they view a concept such as Composable Architecture after learning about NS.

Between September 2024 and November 2024, the Master in Enterprise IT Architecture (MEITA) students of the Antwerp Management School were exposed to NS theory for sixteen hours. The MEITA is an executive master, and students in the program already have many years of practical experience in IT Architecture. At the end of this period, they were given the Gartner paper on the CAF and asked to read it using their newly acquired knowledge of NS. All students received the reading material beforehand, were asked to read it, discuss it in groups for 30 minutes, and then provide a summary of their findings in 5 minutes. In total, 18 students participated, split into four groups.

The first group had difficulty understanding the meaning of PBCs. They realized that the approach can only succeed if business and IT people define the PBCs. They expected more guidance on how to do this and foresaw evolvability issues, as no mechanisms were foreseen to adequately address SoC beyond business changes (what about technological changes?).

The second group tried to list the pros and cons of the framework. They found the use of no- or low-code with a marketplace of PBCs to be powerful. They considered the recommended use of headless promotion technologies to be a good way to decouple UI and logic, while adhering to SoC at least for those two concerns. Group one also struggled with understanding PBCs and claimed they had never observed it in their practices. They see the framework's application more in classifying your existing applications according to business

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

capabilities, and then use the Composable Architecture Integration platform to recombine existing applications.

The third group also struggled to understand PBCs. They compared the main characteristics of PBCs with the NS theorems. Modularity resonates with NS, but the framework struggles to describe the level of granularity. They see SoC applied to some extent but insufficient to be NS compliant. They saw that the second characteristic, autonomy, would require SoS, but this is not mentioned. They make a similar remark about PBC orchestration. They conclude that the fourth characteristic, discoverability, fails to account for the need for AvT and DvT.

The fourth group noticed that while NS is about change over time and minimizing ripple effects, PBC is about building fast. The characteristics of PBC are such that nobody in their right mind would want them, but the framework is insufficiently explained on how to do it. On the other hand, NS tells you how to do it, and doing it the NS way requires significant effort. The group also identifies the PBC granularity uncertainty as a source of future issues. For instance, the sales business capability, packaged in a PBC, may or may not answer to the different sales business capabilities needs in large and complex organizations, opening the door for violation of the four NS principles quicker than expected.

The different groups struggle with similar issues. Once exposed to NS, one can ask more profound questions about how implementation should occur and whether a proposed solution has the characteristics it claims to have. It is important to note that this does not necessarily mean that NS is the optimal solution to operationalize the concept of PBC, as only a single methodology was evaluated in the focus group. Nevertheless, it does seem to validate that there is a need for a strategy to operationalize a PBC architecture through more concrete guidance, such as provided to a certain extent by NS.

The focus group reports align with our findings from the additional literature review, as discussed in Section IV. As CAF becomes more widely known, practitioners start to notice the gray areas and try to fill them with heuristic knowledge that often is a practical application of the NS concepts. NS never claimed to have found something new, but it does claim to have a theoretical and scientific method for unifying and grounding existing heuristic knowledge.

The same experiment was conducted in November 2025 with the new MEITA student cohort. Three groups of three students were formed and were asked to analyse CAF with NS. The reported issues are similar to those found by the previous cohort.

VII. Extending the CAF

In the previous section, we pointed out several shortcomings of the CAF. We will now try to guide you on how to overcome these shortcomings. Wherever possible, we will provide direction supported by science, though this will not always be possible. We will start with guidance on the definition of business capabilities, followed by the size of PBC modules. We continue with PBC Types guidance, address CCCs, governance of the PBC platform, and end with an overview of our advice.

A. Defining PBCs

Before we can package the business capabilities, we need to define them. In Section V-A, we have seen that a common understanding of what a business capability is and how it should be mapped remains lacking.

Innocom, an enterprise architecture consultancy and education institution, offers a two-day seminar on PBCs. In their approach, participants are asked to identify the essential tasks in a given context — the top tasks. They refer to the book ”Top Tasks: A how-to guide” by G. McGovern [30], that helps distinguish top tasks from small ones. Those top tasks are good candidates for being packaged. If those top tasks are considered your top required business capabilities, listing them will provide you with the required PBCs.

We have seen in Section V-A that THE functional view, and thus the capabilities view, of an enterprise does not exist. As such, we might as well try to find one that is as close to reality as possible using the top tasks method. The results of such a bottom-up approach may be validated/adjusted/augmented by top-down approaches. There may exist a publicly/privately available business capabilities map that can serve as validation or a challenging framework for your top tasks. The company may have a strategic map linking mission, vision, strategy, and required capabilities. Both types of business capability maps risk being abstract and disconnected from the nomenclature used by business people in their day-to-day work. Abstraction is suitable for general usability and re-use, but it can hamper adoption and buy-in. We recommend finding a middle ground that connects the abstract and the concrete, and using anthropomorphic naming for the identified business capabilities to ensure maximum unambiguous understanding.

One could use an advanced enterprise engineering method, such as DEMO [22], that is backed by science, to construct the enterprise ontology and, from there, enforce a one-to-one mapping to a set of enterprise functions/capabilities provided by the identified ontological transactions. The ontological transactions could be used as a business capability model because they are outcome-focused (a key feature according to some definitions of business capabilities (see Section V-A)).

Gartner recommends forming fusion teams to ensure both business and IT are at the table and to foster a common understanding. This is good advice. We recommend adding people with a strong methodological background who can distinguish and steer discussions away from technological implementation, and who can reformulate and generalize concepts raised during the interactive workshops among all relevant parties. Such profiles will help in the convergence between bottom-up and top-down approaches.

B. PBCs modularity and granularity

PBCs are considered modules. The raison d’être of PBC is to have a composable architecture that allows combining,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

recombining, and changing PBC to create new applications. The PBCs must favour evolvability.

As NS is a theory of modularity, it provides guidance on how to create and size your modules to enable evolvability. Recall that a system is considered evolvable if it is free of combinatorial effects. A combinatorial effect (CE) occurs when a bounded functional change results in a mount of work that is proportional to the size of the system. The necessary condition for a modular structure to be free of CE is given in the 4 NS theorems. In what follows, we will translate NS into the PBC context and demonstrate the need for compliance through an example.

1) Separation of Concern: The first NS theorem—Separation of Concerns (SoC)—states that a module should encapsulate only one concern, where a concern is defined as a change driver. As such, a PBC should encapsulate only one concern. If multiple concerns are combined in a PBC, than those concerns would need to change all together as a group and not individually, and that would basically mean that the change driver can only exist at the PBC level, not in its components. Our advice is two-fold: either you create PBCs that respect SoC, or you do not allow variations in a PBC.

Let us elaborate by means of an example. The ”invoicing” business capability is part of every company. It makes sense to create a PBC for invoicing that is used across the company to receive payment for delivered services. But what do we mean by invoicing? Is it only the creation of the invoice in electronic format that states the amount to be paid, or do we also include sending the invoice via e-mail, an e-invoicing system, or physical printing and sending? Invoicing risks to contain multiple concerns. NS tells us to split the PBC into smaller PBCs: invoice_calculation, invoice_e-sending, invoice_e-mailing, invoice_printing, invoice_physical-sending. If the PBC is not divided into these smaller modules, CE will be introduced in the landscape. Let us assume that a company offer five services, each delivered by a separate branch. Suppose they all use the PBC ”invoicing”. One of the branches uses the PBC invoicing to calculate and e-send the invoice, another only to calculate and print. Different branches use different concerns packed in the same PBC. Suppose now that a new version of e-sending is required (that may be country-specific). But not all branches require this latest version, and by introducing a bounded functional change, we impact all parts of the organization that use the PBC, meaning we introduce a CE. One could argue that this could be solved by integrating into the PBC the two different versions of e-sending, where the original one stays untouched, and the new one is added, thus being compliant with the Open/Close principle of Robert Martin: modules are open for extension, closed for modification [31]. Suppose now that the company acquires another company to be integrated as a sixth branch, and it needs to set up its invoicing with the invoicing PBC. As the PBC is a package that hides its insides, how will it know which e-invoicing procedure to use? The insides of the PBC need to be known to allow implementation. This contradicts the idea of a module acting as a black box and the concepts of packaging. Combining multiple concerns into the invoicing PBC would mean that all branches need to use the ”invoicing” PBC and will use it in the same way: first calculate, then e-mail, then e-invoice, then print, and finally physically send. You use the full functionality of what is defined as invoicing, or you cannot use it. The moment there is a need for a different use of that PBC, there are various change drivers, and they must be split off. By splitting, we really mean the creation of a PBC to address that concern (like PBC ”e-sending_europe and ”e-sending_india”), not create a new version of the original PBC ”Invoicing V2” that will include the new version of e-sending for India.

2) Separation of State: The second NS theorem is Separation of State (SoS). It postulates that each module should maintain its own state and expose it. As PBCs are modules, they should also preserve and expose their state. Fine-grained PBCs enable more detailed state tracking and actions based on those states. While coarse-grained PBCs hide the orchestration of their inner workings from the external world, fine-grained PBCs require an externally visible orchestration PBC to allow partial execution of the smaller PBCs.

Besides the default requirement that a PBC maintain state, we advise treating the PBC as an atomic transaction and asking whether it can be treated that way. If not, splitting the PBC becomes a necessity.

We would also like to warn of potential Babylonia confusion of the word ”state”. At the technology level, the default standard is stateless APIs. PBCs will be accessible via APIs. As such, some may think that the need for stateful PBCs is opposed to the usage of stateless APIs. The meaning of ”state” is different and has little to do with the other. When working with fusion teams, it might be necessary to skip the discussion of the significance of the word ”state” asap to avoid further misalignment.

In our invoicing PBC example, the module should have states as ”invoice being calculated”, ”invoice calculation error - error description: xxx”. In the previous subsection, we advised either using a single PBC or splitting it into fine-grained PBCs. When the PBC is kept as a whole, a grouping of tasks and data, it should expose state about that whole rather than its inner workings. Issues about the inner workings should be part of logging, of course, for troubleshooting purposes, but as the inner tasks of the PBC are not allowed to be used, they cannot be re-called or acted upon. If the coarse-grained invoicing PBC had a state of ”error sending e-invoice”, the entire ”invoicing” transaction must be rolled back. If we had fine-grained PBCs, the invoice_calculation, invoice_e-mailing, invoice_printing and invoicing_physical-sending could be ok, but the invoice_e-sending could have failed, only requiring a rollback of that transaction. This should make it clear that going for the coarse-grained PBC, if SoC is ignored, SoS gives you an additional sanity check - are you really sure you want to do this, because you will now have to solve rollback issues, complexifying the PBC.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

3) Action Version Transparency: The third NS theorem is Action version Transparency (AvT). It postulates that an internal change to the implementation of a module, assuming iso functionality, should not impact the calling module.

Applying this theorem to PBC is straightforward: all available interfaces to the PBC must remain the same when internal upgrades to the PBC are performed. Examples of internal upgrades that should not affect the interface include performance improvements, bug fixes, and changes in the underlying technology. To be even more precise, the API URLs should not change.

Changes in functionality can break the interface. If this is the case, the PBC should no longer remain the same PBC; a new one must be created. The new PBC should have a different anthropomorphic name that reflects its new function and exist alongside the previous PBC. Users of the old functionality are not affected, and those requiring the new functionality connect to the new PBC. Suppose the first functionality is to be replaced over time. In that case, an orchestrated transition from the old to the new can occur—a gradual migration to the new URL, potentially facilitated by a platform rejuvenation facility (see further).

Our advice regarding AvT is to adopt a clear versioning strategy that distinguishes between internal and external versioning. Internal versioning should not impact the PBC; external versioning is subject to migration effort that can be proportional to the system's size, but this can be resolved through automation (see further).

Let's apply this to our Invoice PBC example. Changes to a new mailing system, a more efficient invoice calculation method, a new printing system, or a new e-invoice system should not affect the functioning of the PBC, nor do they justify a new version (with a new name). Only the internal versioning is updated. Suppose a new e-invoicing system has a different interface from the current one, and that this change breaks the interface. The internal change ripples through to the PBC. At this point, we have identified a change driver, and Soc tells us we should split off this module. Maintaining AvT also pushed toward finer granularity. If we do not want to create a new separate PBC dedicated to e-invoicing, then we must create a new version of the invoicing PBC. That new version should have a new name, like "invoicing with new e-invoicing," to indicate the functionality difference from the original. Note that such a solution will require maintaining two PBC that share the same code for invoicing_calculate, invoicing_print, invoicing_e-mail, and invoice_physical-sending. At this point, we start introducing the ripple effect and even CE when future changes to those PBC internals are required. Unless the PBCs are an aggregation of more fine-grained PBCs that the business cannot access (see further).

4) Data Version Transparency: The fourth and last NS theorem is Data version Transparency (DvT). It postulates that adding information to the data structure passed to a processing function should not affect the processing function.

PBCs will perform actions on data (see Section VII-C) that will be passed to them via APIs or messages. The data transmission architecture or directive must favour DvT. This means that in a distributed environment, we favour sending entire objects (packed in JSON or XML, for instance) (aka stamp coupling) rather than passing individual attributes (aka data coupling). In an in-memory setup, this means objects are passed by reference, not by value. Security professionals will typically oppose this, as disclosing too much information may lead to misuse or leakage. However, if the PBC are adequately implemented, with security as being part of their design, then this should not be a problem as there are sufficient methods available to shield, anonymize and validate data access authorization. This does imply that security must be set at a fine-grained level. It is beyond the scope of this paper to discuss data security, but from an evolvability perspective, pass-by-reference is superior to pass-by-value.

Our advice regarding DvT is to ensure a data transmission architecture is put in place that supports secure stamp coupling for PBC-to-PBC communication.

C. PBC Types

According to the CAF, PBCs are of a specific type. Gartner is unclear whether they mean there should be a one-to-one mapping between a PBC and a PBC type, or whether they suggest that a PBC should be built with sub-modules of a specific type. Those sub-modules are no longer PBCs, they do not address a business capability but rather a generic task or thing.

NS introduces the notion of elements, the elementary building blocks an application is made off. Whether an application represents the implementation of a single business capability or a collection of business capabilities working together, it will be built using elements. Elements define the basic operations performed by an IT system: represent data, perform tasks on data, execute tasks in a specified order (orchestration/choreography), connect to other elements, and trigger the actions an element is supposed to perform. In our opinion, Gartner makes the PBC approach overly complicated by introducing PBC types. It suffices to say that PBC should be made out of smaller modules of a specific type/pattern/element. As those types/patterns/elements are generic by nature, they no longer represent a particular business capability.

Our advice is to drop the PBC Types and to see PBCs as a composition of generic data, action, flow, trigger and connection patterns/elements. Those patterns/elements should be highly standardised and also be compliant with the NS principles. We will return to this point in the section on PBC platform governance.

In our example, what type of PBC should we assign to our invoice PBC? The PBC will contain data and the actions taken on it. If the PBC then Data PBC Type or an Application Action PBC Type? And if it is the combination of the two, does that then mean that the invoice PBC is a combination of

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

an invoice Data PBC type and a calculate invoice Application Action PBC Type, etc.? These questions move us away from what is essential (the functionality of the PBC). They can be shortcut by treating all PBC as composed of data, action, flow, trigger, and connection patterns/elements.

D. PBC and CCCCs

In the previous section, we considered PBCs as composed of elementary patterns implemented in technology. To protect against technological evolution, that cuts across all implementations, we need to ensure that technological concerns or change drivers are adequately separated from the core elementary functionality. Elementary patterns need to be carefully designed for this. If not, technological changes will impact all patterns, PBCs and PBC instantiations.

There can also be CCC at the PBC level. A classic example is security. Our invoicing PBC needs, by design, to include security. Only those authorized to use the PBC should be allowed to do so. All PBCs must incorporate this concern by design. This will result in each PBC having some elementary patterns associated with it to take care of authorization. And those elementary patterns must have the technological aspect of authorization properly separated.

Our advice is to look for CCC at the PBC level, and security must be there by default. If not, the security-by-design or zero-trust principle erodes from the top.

E. PBC Platform Governance

According to Gartner's CAF, the PBC platform should handle discovery, composition (via low-code/no-code), and deployment of PBC. We will handle them one by one.

1) Discovery: In previous sections, we motivated the usage of anthropomorphic naming for the PBCs. Next to naming the PBCs, there should also be a hierarchy of PBCs that allow the creation of a name space. From public and privately available business capability frameworks, we know that business capabilities are organized in a hierarchy. We advise using an anthropomorphic hierarchical name space to classify business capabilities and PBCs.

Returning to our invoicing example. Within the company of one of the authors, invoicing is present across multiple business capabilities - see Figure 4. There are three invoicing business capabilities: one for certificates, one for commodities and one for services. Note that those BCs aggregate billing and invoicing. Although related, they differ in the actions required. If we follow our own recommendations, billing and invoicing should be split at the second level of the Sales Management BC into "Billing Management" and "Invoice Management", and level three should be split accordingly (see Figure 5). We could now have three PBCs: one for certificate invoicing, one for commodity invoicing, and one for service invoicing, or we could assign to those three BCs, only one PBC, being our invoicing PBC. From a Business Capabilities Framework perspective, it makes sense to state that we bill for three reasons clearly. Still, at the implementation level, a single set of tasks must be performed. This leads us to the advice that the Business Capabilities Framework must be part of the PBC platform and should serve as a path toward a PBC, but not as the list of PBCs themselves. Some BCs might not even require a PBC.

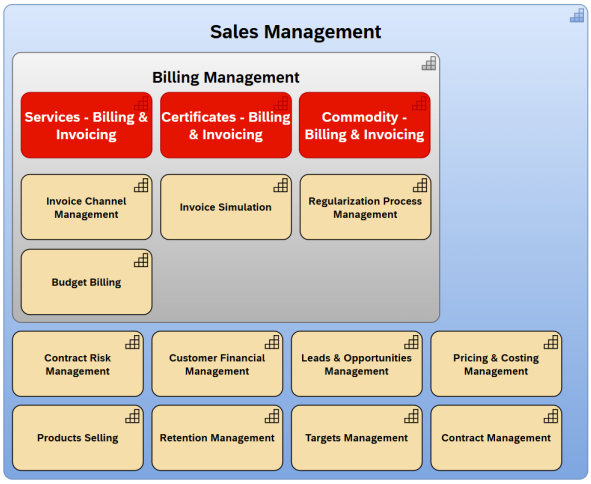


Fig. 4. Business Capabilities Framework - invoicing

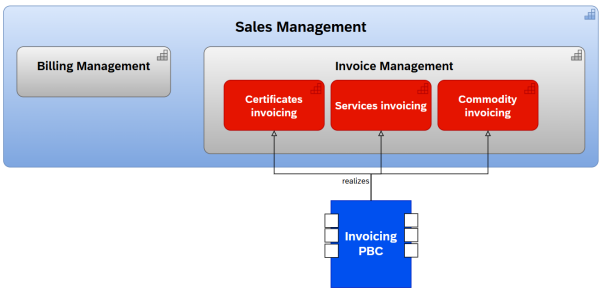


Fig. 5. Linking the Business Capabilities Framework to PBCs

2) Composition: To facilitate the reconstruction of PBCs in Composed Applications (CA), Gartners proposes a low-code/no-code platform. Although low-code/No-code is in line with McIllroys dream, we warn for the vendor/technology lock-in they represent. Migrating from one low-no-code platform to another will be a combinatorial effect. As those platforms are not open source, it isn't easy to assess whether they respect NS principles. We advise looking for and selecting low-code/no-code platform constructs that allow breaking PBCs into elements (data, task, trigger, flow, and connect), make PBCs with them, combine the PBCs into CAs and governing the platform accordingly. The platform may offer you appealing packaged features, but these may introduce CEs. Technological features often lure you into using them. It's not because something is technically possible or the technology is available to do so that one must do it. We are even more in favour of working according to the NS development

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

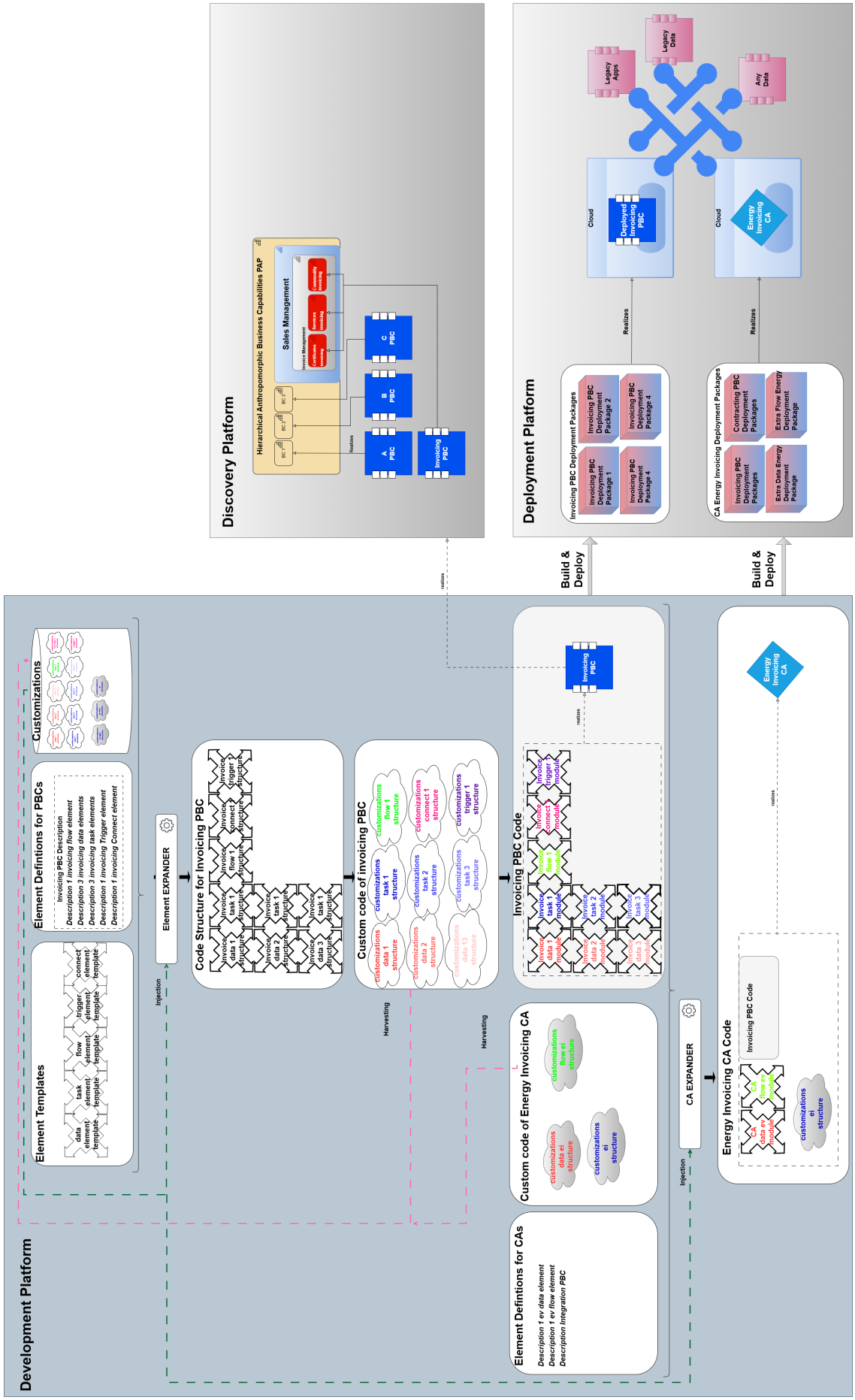


Fig. 6. Proposed PBC platform.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE III
RECOMMENDATIONS TO IMPROVE GARTNER'S CAF

Area of Improvement	Recommendation Summary	Key Actions / Guidance
Defining Business Capabilities	Use a combination of bottom-up (top tasks) and top-down (strategic maps, frameworks) approaches. Employ anthropomorphic naming for clarity and adoption.	• Identify essential tasks ("top tasks") as candidate business capabilities. • Validate/augment with strategic maps or industry frameworks. • Use DEMO to identify essential transactions and capabilities. • Use names that reflect real-world business language for better understanding.
Granularity of the Modular PBCs	Use NS evolvability theorems	• Apply SoC, SoS, AvT and Dvt
Separation of Convern (SoC)	Ensure each Packaged Business Capability (PBC) encapsulates only one concern (change driver). Split PBCs as needed to avoid combinatorial effects.	• Split PBCs until each addresses a single concern. • Avoid variations within a PBC; create new PBCs for different usages. • Treat PBCs as atomic transactions when possible.
Separation of State (5oS)	Each PBC should maintain and expose its own state. Fine-grained PBCs enable better state tracking and rollback.	• Design PBCs to track and expose relevant states. • Prefer fine-grained PBCs for detailed state management. • Clarify "state" meaning to avoid confusion between business and technical contexts.
Action Version Transparency (AvT)	Internal changes to a PBC should not affect its interface. Use clear versioning strategies.	• Maintain stable APIs for PBCs. • Distinguish between internal and external versioning. • Create new PBCs (with new names) for breaking changes, enabling gradual migration.
Data Version Transparency (DvT)	Data transmission between PBCs should favor "stamp coupling" (passing whole objects) for evolvability.	• Implement secure, object-based data exchange (e.g., JSON/XML). • Ensure security at a fine-grained level. • Prefer pass-by-reference over pass-by-value for evolvability.
PBC Types	Drop rigid PBC types; instead, compose PBCs from standardized, generic elements (data, action, flow, trigger, connector) compliant with NS principles.	• Build PBCs from NS-compliant elements. • Avoid unnecessary complexity from PBC type distinctions. • Standardize patterns/elements for reusability.
CCCs	Identify and modularize CCCs (e.g., security) at both the PBC and element level. Ensure technological concerns are separated from core functionality.	• Integrate CCCs (like security) by design in every PBC. • Separate technological aspects from business logic. • Use elementary patterns for CCCs.
PBC Platform Governance	Enhance platform features for discovery, composition, deployment, and data integration. Apply NS principles to platform design.	• Use hierarchical, anthropomorphic naming for PBCs. • Link business capability frameworks to PBCs. • Select low-code/no-code platforms that support NS elements. • Apply NS principles to deployment and data integration. • Avoid vendor lock-in and ensure platform evolvability.

methodology. The basis for this are the NS elements, that are made up of code templates that can be instantiated into boilerplate code via expansion. To decide what data, task, flow, trigger, and connection elements you require, we recommend starting by creating a data model tailored to the PBC building blocks you need. This does not differ from Evans' advice in Domain-Driven Design (DDD) [32]: start by creating a bounded context. Next to the data, the task that will work on that data and the orchestration of those tasks. For this, state-machine models are recommended. The boilerplate code can now be customised, and those customisations can be separated from the boilerplate using harvesting, allowing them to be reinjected when the templates are updated—the rejuvenation process. The resulting library of PBCs is to be linked to the discovery features — the list of BCs and their corresponding PBCs. To create CAs, a set of PBCs and elements that facilitate integration and orchestration will need to be expanded and packaged. Again, a data model and a state machine can be used to represent the CA's logic. We propose a CA-expander that integrates, orchestrates (flows), and connects to PBCs.

3) Deployment: Thanks to build frameworks, the cloud, and Infrastructure as Code, deployment has become a less

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

painful process than in the past. The application of deployment pipelines and the organisation of teams according to the DevOps philosophy [33] is now a standard in the organisation and implementation of IT Operations. We want to point out that, of course, building your own deployment infrastructure or using a platform for it is, in itself, a vital change driver, and again, NS principles should be applied to facilitate deployment evolvability. We refer to [34] for more information on the subject.

4) Data Fabric: Regarding the Data Fabric component of the CAF, we give the same advice as for the PBC Types: ignore it. When the PBC needs data it can tap into that data via data elements. The data elements ensure proper encapsulation. Any data-providing platform, be it a simple DB or an advanced/abstract data Fabric/Lake/Lakehouse (whatever term is used), should respect the NS principles so that changes do not ripple through to the PBCs. If this cannot be guaranteed by the data provided, the data should not be integrated in the PBC, as they will hinder evolvability

F. Overview

Table III provides an overview of the recommendations we have put forward in the previous sections. Figure 6 gives a graphical overview of our proposed enhanced CAF. On the left side of Figure 6 you will notice the ”Development Platform”. It contains the different ”Element Templates” (data, task, flow, trigger, connection, and other template types can be added) and the ”Element Definitions for a PBC”, containing the parameterization of the element templates. The combination of both goes into the ”Element Expander” and results in the code structures (the boilerplate) code of a PBC. These code structures are customized and all customizations are harvested and put in the ”Customizations” repository. During rejuventation (updates of the ”Element templates”), the ”Element Expander” will inject the customizations in the boilerplate code. The result is the PBC code. CAs are a combination of PBCs and some glue or custom coding. This is shown at the bottom left, where the ”CA Expander” combines extra elements and existing PBC code into CAs. Again, all customizations are harvested, and reinjection can happen during rejuvenation. To the right, you will note the ”Discovery Platform”, that will link the available PBCs to the BC framework. At the bottom right, the ”Deployment Platform” is shown. It contains the tooling to package the PBCs and CAs code into deployable units that end up in the cloud, where they can interact with all other applications (PBCs, CAs, and legacy).

VIII. Conclusion

Application re-use is not just a long-forgotten dream of McIllroy, but a focus of many companies. The ability to reuse and recombine applications to support the changing business conditions is an expectation many CEOs have of their CIOs. In recent years, the focus on functional reuse has been pushed aside by technology-focused integration patterns. Gartner puts functional re-use back on the map with its CAF, where PBCs are the essential building blocks of application landscapes. By using NS as an instrument of design and evaluation method for evolvable systems, we pointed out operationalization issues one might face when trying to implement the CAF, or one might use it to evaluate providers of solutions based on it critically. A focus group was used to validate and balance our findings. In this extended edition of our original paper, we convert our criticism into recommendations and improvements to the CAF, combining knowledge of NS and some heuristics, to arrive at a proven CAF for evolvable PBCs.

Acknowledgment

The authors thank Rudy Claes of Innocom for introducing them to Gartner’s Composable Architecture Framework and conducting a brainstorming session about the framework and its evolvability. We also thank the Master in Enterprise IT Architecture (MEITA) students at Antwerp Management School (AMS - cohort 2024 and 2025) for their contribution as focus group.

References

[1] G. Haerens and H. Mannaert, ”On the operationalization of composable architecture by means of NS theory,” In PATTERNS 2025: The Seventeenth International Conference on Pervasive Patterns and Applications, pp 23-30, 2025.

[2] G. Coulouris, J. Dollimore, and T. Kindberg, ”Distributed Systems: Concepts and Design Edition 3,” ISBN:978-0-201-61918-8, 2001.

[3] J. Sun and Y. Natis. ”Use Gartner’s Reference Model to Deliver Intelligent Composable Business Applications,” Gartner, ID G00720701, 2020 - refresh 2022.

[4] H. Mannaert, P. De Bruyn, and J. Verelst, “On the interconnection of cross-cutting concerns within hierarchical modular architectures,” IEEE Transactions on Engineering Management, Vol. 69, pp. 3276-3291 2020.

[5] Y. Natis and G. Alvarez, ”How to Implement Composable Technology with PBCs,” Gartner, ID G00751018, 2021.

[6] H. Mannaert, J. Verelst, and P. De Bruyn, “Normalized Systems Theory: From Foundations for Evolvable Software Toward a General Theory for Evolvable Design,” ISBN 978-90-77160-09-1, Koppa, 2016.

[7] H. Mannaert, J. Verelst, and K. Ven, “The transformation of requirements into software primitives: Studying evolvability based on systems theoretic stability,” Science of Computer Programming, Volume 76, Issue 12, pp. 1210-1222, 2011.

[8] P. Huysmans, J. Verelst, H. Mannaert, and A. Oost, ”Integrating information systems using normalized systems theory: four case studies,” In IEEE 17th Conference on Business Informatics, Volume 1, pp. 173-180, 2015.

[9] A.W. Scheer, ”The Composable Enterprise: Agile, Flexible, Innovative: A Gamechanger for Organisations, Digitisation and Business Software,” ISBN:978-3-658-42482-4, Springer, 2024.

[10] I. Ivas, ”Implementation of Composable Enterprise in an Evolutionary Way through Holistic Business-IT Delivery of Business Initiatives,” In Proceedings of the 26th International Conference on Enterprise Information Systems, Volume 1, ISBN: 978-989-758-692-7, pp. 397-408, 2024.

[11] MACH Alliance, [Online], Available: https://machalliance.org, [retrieved: November, 2025].

[12] Gartner DXP Magic Quadrant, Available:https://www.gartner.com/reviews/market/digital-experience-platforms, [Online], [retrieved: November, 2025].

[13] MACH Alliance, [Online], Available: https://machalliance.org/insights-hub/composable-comes-of-age-in-the-gartner-dxp-magic-quadrant, [retrieved: November, 2025].

[14] ESICONF2025, [Online], Available: https://www.peernetwork.it/en/esiconf2025-agenda/, [retrieved: November, 2025].

[15] Global Logic Practitioner Perspective, Available: https://www.globallogic.com/wp-content/uploads/2023/04/Composable-Enterprise-PBC.pdf, [Online], [retrieved: November, 2025].

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[16] Forbes, Available: https://www.forbes.com/councils/forbestechcouncil/2025/07/22/20-hidden-benefits-of-composable-architecture-in-enterprise-tech/, [Online], [retrieved: November, 2025].
[17] Contentstack, Available: https://www.contentstack.com/blog/composable/why-your-business-needs-a-packaged-business-capability-now, [Online], [retrieved: November, 2025].
[18] Value Innovation Labs, Available: https://valueinnovationlabs.com/blog/digital-transformation/why-composable-architecture-will-lead-digital-transformation-in-2025/, [Online], [retrieved: November, 2025].
[19] HCLSoftware, Available: https://www.hcl-software.com/blog/commerce/embracing-packaged-business-capabilities-pbcs-a-superior-middle-alternative-to-microservices-vs-monoliths, [Online], [retrieved: November, 2025].
[20] J. Barney, ”Firm resources and sustained competitive advantage,” In Journal of Management, Volume 17, Issue 1, pp 99-120, 1991.
[21] T. Offerman, C.J. Stettina, and A. Plaat, ”Business capabilities: A systematic literature review and a research agenda,” In International Conference on Engineering, Technology and Innovation (ICE/ITMC), pp. 383-393, 2017.
[22] J.L.G Dietz and H.B.F. Mulder, ”Enterprise ontology: A human-centric approach to understanding the essence of organisation”, ISBN:978-3-031-53361-7, Heidelberg: Springer, 2024.
[23] BIAN, [Online], Available: https://bian.org, [retrieved: November, 2025].
[24] NBility, [Online], Available: https://www.edsn.nl/nbility-model, [retrieved: November, 2025].
[25] LeadingPractise, [Online], Available: https://www.leadingpractice.com, [retrieved: November, 2025].
[26] SAFe Framework, [Online], Available: www.scaledagileframework.com, [retrieved: November, 2025].
[27] Object Management Group (OMG), [Online], Available: https://www.objectmanagementgroup.org/, [retrieved: November, 2025].
[28] The Open Group Architecture Framework (TOGAF), [Online], Available: https://www.opengroup.org/togaf, [retrieved: November, 2025].
[29] BIZBOK, [Online], Available: https://www.businessarchitectureguild.org, [retrieved: November, 2025].
[30] G. McGovern, ”Top Tasks – A how-to guide”, ISBN: 978-1-9164446-1-4, Silver Beach, 2018.
[31] R.C. Martin, ”Clean architecture: a craftsman’s guide to software structure and design”, ISBN-13: 978-0-13-449416-6, Prentice Hall Press, 2017.
[32] E. Evans, ”Domain-driven design: tackling complexity in the heart of software”, ISBN-10: 0321125215, Addison-Wesley Professional, 2004.
[33] K. Gene, et al, ”The DevOps handbook: How to create world-class agility, reliability, & security in technology organizations”, ISBN: 978-1942788003, IT Revolution, 2016
[34] H. Mannaert, T. Van Waes, and F. Hannes, ”Toward a rejuvenation factory for software landscapes”, Patterns 2024: The Sixteenth International Conference on Pervasive Patterns and Applications, April 14-18, 2024, Venice, Italy, 2024.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Enabling Kubernetes Workload Execution on Rootless HPC Systems with KSI: A Slurm Integration Framework

Jonathan Decker*, Mojtaba Akbari†, Ali Doosthosseini*,
Sören Metje*, Aasish Kumar Sharma† and Julian Kunkel*
*Institute for Computer Science, University of Göttingen,
Goldschmidtstraße 7, 37077 Göttingen, Germany
e-mail: jonathan.decker@uni-goettingen.de | adoosth@gwdg.de
soerenmetje@yahoo.de | julian.kunkel@gwdg.de
†Working Group Computing, GWDG,
Burckhardtweg 4, 37077 Göttingen, Germany
e-mail: mojtaba.akbari@gwdg.de | aasish.sharma@uni-goettingen.de

***Abstract*—Kubernetes has become a widespread orchestrator for cloud workloads but with increasing demand for compute the need arises to also access HPC environments that are operated via batch schedulers such as Slurm. A number of solutions for combining Slurm and Kubernetes are available, which can be categorized further based on the interaction between Slurm and Kubernetes that they provide. We consider the use case of utilizing an existing Slurm cluster to run Kubernetes workloads. In a previous publication we had introduced Kind Slurm Integration (KSI) based on Kind and rootless Podman and compared its performance, usability and maintainability to the existing solutions Bridge Operator and High-Performance Kubernetes (HPK). We found that Bridge Operator provides native performance as it effectively submits Slurm jobs through a Kubernetes interface and that HPK provides good performance by creating almost feature complete Kubernetes clusters on top of Apptainer. KSI on the other hand was able to provide fully functional Kubernetes clusters inside Slurm jobs but lacked behind in network performance and did not support multi-node clusters. In this work we present a new version of KSI with improved network performance through bypass4netns and support for multi-node clusters via Liqo. Overall, we conclude that running Kubernetes workloads under Slurm is possible with acceptable overhead in terms of performance and without missing out on features.**

Keywords-Kubernetes; HPC; Container; Slurm; Cloud.

I. Introduction

This work discusses an extension of our previous publication [1], which had introduced the first version of KSI. We have since then improved on it and addressed several shortcomings. The motivation for the development of the first version and the improved version of KSI are the same and are given in the following.

Kubernetes has established itself as a widespread solution for orchestration of cloud workloads [2][3] and is used for various workloads including service computing, running large amounts of micro services, as well as batch jobs, such as data analytics or machine learning. However, batch jobs would fit better into HPC environments where powerful high-performance compute and networking resources are available. HPC workloads are commonly scheduled using a batch scheduler such as Slurm [4] but Kubernetes itself can also be used for scheduling HPC jobs using a batch scheduler such as Volcano [5] and have already been scaled to large clusters using appropriate workarounds [6]. Nevertheless, while Kubernetes brings a large array of features, its virtualization layers incur a performance overhead compared to the bare metal performance [7] that could be achieved with Slurm.

Users might want to bring their Kubernetes workloads into Slurm-based HPC environments to benefit from the reduced overhead compared to a regular Kubernetes cluster or to gain access to additional compute hardware, which could also include specialized hardware only available in HPC environments. However, rewriting Kubernetes workloads to be executable in Slurm might require significant effort and expertise with the scheduling systems. Nevertheless, various approaches and implementations exist for combining Slurm and Kubernetes enabling users to dynamically move workloads between cloud and HPC environments.

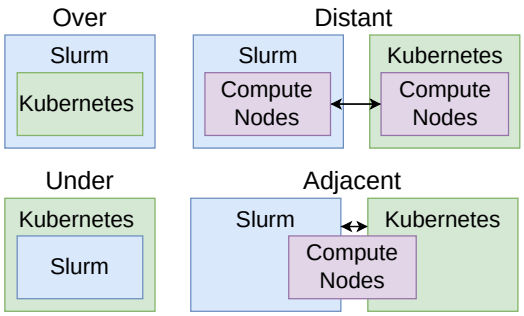


Figure 1. Overview of the four integration models based on the definition by Wickberg [8]. In the **Distant** model, the arrows represent a possible system for negotiating resources between Slurm and Kubernetes such that nodes could be moved as needed. In the **Adjacent** model, the arrows represent bridging tools that enable the execution of workloads under the other scheduling system, respectively.

As there have been various efforts to combine Kubernetes and Slurm, we consider the definition by Wickberg of Schedmd [8] who defines four categories from the perspective of Slurm for such approaches.

- **Over**: The entire Kubernetes environment exists within a Slurm job and is therefore temporary as it is fully removed once the job completes.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

- **Distant**: Compute nodes are part of either a Kubernetes or a Slurm cluster and may be moved between the clusters.
- **Adjacent**: Slurm and Kubernetes utilize some form of plugins or bridging tools to cooperate but can still be used individually.
- **Under**: Kubernetes runs a Slurm cluster within its own environment across one or more pods.

These four models are visualized in Figure 1.

Given the above use case of running Kubernetes jobs in an existing Slurm environment, this fits the **Over** or **Adjacent** model. We have investigated existing solutions that implement either of these models and found various approaches that provide the **Adjacent** model but no system for having a Kubernetes cluster running within a Slurm job as described in the **Over** model. Therefore, we present **Kind Slurm Integration (KSI)** [9], an implementation of the **Over** model based on Kubernetes in Docker (Kind) [10]. We have systematically evaluated and compared KSI to existing solutions including **Bridge Operator** by IBM [11], **WLM-Operator** by Sylabs [12], **kube-slurm** by Kalen Peterson [13] and **High-Performance Kubernetes (HPK)** [14].

Our evaluation consists of a review of the state of the respective projects with regard to features and maintainability as well as a performance analysis to determine the overhead incurred by the respective approach. For this purpose, we benchmarked the solutions based on workload startup time, CPU compute performance, memory throughput, storage throughput, network latency as well as network throughput, and compared the results to bare metal. We found that not all of the implementations listed above were able to pass a minimal functionality test. For those that passed, no significant differences in CPU compute performance, memory throughput and storage throughput were found. However, our original implementation of KSI, as presented in [1], was outperformed in terms of startup time and network performance.

Furthermore, we had investigated whether a given approach supports all Kubernetes features including controlling workloads through `kubectl` and providing Kubernetes network abstractions such as services. We found that KSI provided the most complete support for Kubernetes features compared to the others. Bridge Operator, for instance, submits Slurm jobs through Kubernetes such that workloads are executed as scripts outside of Kubernetes. In the case of HPK, it treats Slurm worker nodes as part of its Kubernetes cluster but does not support `kubectl exec` and Kubernetes services.

In this paper we present an improved version of KSI with two additions that enable it to overcome its most significant shortcomings. To ensure it is clear what version of KSI we are referring to, we will note the original version of KSI as introduced in [1] as KSI 1 and the new version presented in this work as KSI 2 for the rest of the paper. If we refer to KSI without specifying which version, then a given statement applies to both versions of KSI unless specified otherwise in the direct context.

KSI 1 presented in [1] did not support multi-node clusters and fell behind in network performance. KSI 2 utilizes bypass4netns [15][16] instead of slirp4netns for rootless container networking, which significantly improves the network performance. Moreover, we have also integrated Liqo [17] such that KSI 2 can now be deployed across all nodes in a given Slurm job and assemble itself into a single Kubernetes cluster. Both versions of KSI implement the **Over** model of combining Slurm and Kubernetes such that it can be executed in a Slurm job without requiring an existing control plane.

Overall, this paper contributes a systematic evaluation of existing approaches that implement the **Adjacent** or **Over** model to combine Slurm and Kubernetes as well as the design and implementation of KSI. We cover both the original implementation in KSI 1 as given in [1] and the improvements introduced in this work for KSI 2. Finally, we provide an overview of the features and limitations of the evaluated approaches. It should be noted that KSI 1 is based on the master's thesis of one of the authors [18].

The remainder of the paper is organized as follows: In Section II, the various implementations for integrating Slurm and Kubernetes are discussed. The methods for benchmarking and comparing the solutions as well as the design of KSI are presented in Section III. The results of the evaluation are given in Section IV. Finally, Section V provides the conclusion and outlook for future work.

II. Related Work

To properly distinguish various approaches for combining Slurm and Kubernetes, we discuss the four models along with notable examples. We also cover related approaches that do not use either Slurm or Kubernetes and then provide a more in-depth look at the implementations, which we evaluated in this paper.

A. Models for Integrating Slurm and Kubernetes

The four categories for combining Slurm and Kubernetes defined by Wickberg of Schedmd [8] are **Over**, **Distant**, **Adjacent** and **Under** as defined in Section I.

*a) **Distant** model:* Notable implementations include [19] and [20], which both implement systems for dynamically changing the partitioning of a node pool between a Kubernetes and Slurm cluster.

*b) **Under** model:* Contributions have been made in [21], [22] and [23], in which Slurm is being run as a set of Kubernetes pods. A significant project in this category is Slinky [24] by Schedmd who had created a Slurm Kubernetes bridge implementation as a proof-of-concept before creating Slinky. The proof-of-concept implementation followed the **Adjacent** model, was not fully functional and has since then been removed from public access.

*c) **Adjacent** model:* Approaches in this category are relatively diverse in their methods including Bridge Operator [11], WLM-Operator [12] and HPK [14]. Each of these approaches is discussed in more detail in Subsection II-C.

*d) **Over** model:* There are no notable implementations of this model except for KSI 1 [25] and KSI 2 [9], which are presented in detail in Subsection III-B.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

B. Other Approaches for Integrating HPC and Cloud

While this work focuses on combining Slurm and Kubernetes, it should be noted that there are alternative approaches to running HPC workloads through a cloud interface. For example, as mentioned in Section I, Volcano [5] is an extension for the Kubernetes scheduler, which implements features such as batch and gang scheduling. This enables the execution of batch workloads through Kubernetes without Slurm as shown in [26][27].

Another notable approach is hpc-connector [28] presented in [20], which enables the submission of jobs through an arbitrary cloud interface to be executed via Slurm. This approach can be considered similar to Bridge Operator but is not bound to Kubernetes but also lacks deeper integration with any specific cloud platform to enable advanced features.

Finally, there is [29], which introduces an integration of TORQUE [30], another HPC batch scheduler, with Kubernetes. This enabled scheduling of HPC workloads through Kubernetes to TORQUE similar to Bridge Operator, which would therefore match the **Adjacent** model.

C. Implementations for **Adjacent** Slurm and Kubernetes

1) WLM-Operator: Sylabs Inc. had developed the WLM-Operator [31] and Singularity-CRI [32] with Singularity-CRI providing a Kubernetes-compatible implementation of the Container Runtime Interface (CRI) for Singularity [12]. The WLM-Operator implements a Kubernetes operator that is able to interface with Slurm such that Slurm nodes become visible in Kubernetes as virtual nodes.

Moreover, it provides a Custom Resource Definition (CRD) in Kubernetes called *SlurmJob*, which enables the submission of Slurm jobs through Kubernetes. When submitting a SlurmJob, a dummy pod is created in Kubernetes and the actual job is submitted to Slurm to be run in a Singularity container. The results are then collected through another pod via a shared storage before closing the dummy pod once the job completes.

However, on December 30th 2020, both WLM-Operator and Singularity-CRI projects have been archived on Github with no further development planned.

2) Bridge Operator: IBM had developed Bridge Operator [33] for a Kubernetes cluster to be able to access external compute resources including Slurm clusters [11]. Bridge Operator implements a Kubernetes operator and provides the *BridgeJob* CRD, which accepts all the details required to launch a Slurm job including the remote URL of a Slurm cluster, what resources to request and a remote storage configuration.

For each BridgeJob, the Bridge Operator starts a monitoring pod and submits the job to Slurm. The monitoring pod regularly updates a Kubernetes ConfigMap with the current status and fetches the job output. The creators of Bridge Operator have also demonstrated how to run Kubeflow workloads through BridgeJobs [34], however, as these jobs are converted to Slurm jobs, no Kubernetes pods are directly being run in Slurm.

3) HPK: HPK [35] is presented in [14] and [36] as a way to run Kubernetes workloads on Slurm through Apptainer [37]. It is deployed as a single Apptainer container that runs the Kubernetes control plane and a custom implementation of virtual Kubelet [38], which presents an entire Slurm cluster as a single node in the cluster. Every time a new pod is to be scheduled, it submits a job through Slurm for the pod to be started as a container using Apptainer.

For the container networking to function, it relies on Flanneld service [39] to be installed on the nodes and the Flannel-CNI plugin [40] to be installed for Apptainer. However, only headless services without cluster IPs are supported as the additional layer of load balancing is not possible with the employed networking stack. Moreover, the command `kubectl exec`, which serves to execute commands inside Kubernetes pods, is not supported.

4) Kube-Slurm: The kube-slurm project [13] provides a tool for controlling Kubernetes resources using Slurm jobs. When deploying, Slurm and Kubernetes must both be installed on the same set of nodes with `kubectl` available on all nodes. Once deployed, users can submit Slurm jobs, which get scheduled by the tool as Kubernetes pods onto the nodes selected by the Slurm scheduler.

The deployment can also be completed with the **Under** model by having Slurm run within Kubernetes but still using Slurm to schedule the pods. Nevertheless, due to the way the access is provided to the Slurm scheduler, all users receive the same access to the Kubernetes cluster making this approach unfit for multi-user setups with potentially malicious users.

III. Methodology

This work focuses on approaches for combining Kubernetes and Slurm that allow for running workloads on an existing Slurm cluster following the **Over** or **Adjacent** model and investigates the suitability of the existing solutions. For that purpose we define the following research questions:

RQ1 Can workloads be submitted using Kubernetes tooling, e.g., `kubectl`?
RQ2 Can workloads be scheduled and executed on machines managed by an existing Slurm cluster without root access?
RQ3 Can workloads be executed across multiple machines in parallel?
RQ4 What is the performance overhead imposed by the tool?
RQ5 Is the tool easy to operate for the end user?
RQ6 Is the tool well maintained?

RQ1, **RQ2** and **RQ3** define the functional requirements. For a solution to be a valid approach for utilizing a Slurm cluster through Kubernetes, it should answer yes to at least **RQ1** and **RQ2** with a yes to **RQ3** being desirable but not strictly required. Notably, these requirements do not include whether a solution must be able to run Kubernetes workloads or if it may run Slurm workloads through a Kubernetes interface. For example, Bridge Operator accepts Slurm workloads submitted through Kubernetes while HPK processes Kubernetes workloads submitted through a Kubernetes interface. Still, both solutions execute the workloads on a Slurm cluster.

Moreover, the distinction between the **Over** and **Adjacent** model breaks down to whether the deployment requires an existing component, such as a Kubernetes control plane. For

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

example, Bridge Operator and HPK, which both follow the **Adjacent** model, require a running Kubernetes control plane, to which a user submits their workloads. The respective tool then submits the workload to Slurm. KSI, on the other hand, implements the **Over** model such that a user would instead submit their workload by calling Slurm.

RQ4 is concerned with the performance cost of a given solution. Depending on the architecture and optimization of a given implementation, it may cost additional compute power or delay the start of workloads, which should be minimal for an application to fully harvest the power of HPC machines.

RQ5 and **RQ6** cover the usability and maintainability of a given software providing an indication for the viability in productive use.

While **RQ1**, **RQ2** and **RQ3** can be answered as yes or no questions, **RQ4**, **RQ5** and **RQ6** require a graded answer. We use a three point scoring from + (good) over ○ (fair) to − (bad) to be able to quickly compare the results for multiple implementations. + is the best score, which is given if the implementation fulfills the requirements without any significant drawbacks. ○ is the middle score, which indicates that some limitations apply and − is the lowest score, which applies if significant shortcomings exist.

A. Selection of Implementations to Evaluate

In Section II-C, we had introduced the WLM-Operator, Bridge Operator, HPK and Kube-Slurm. Before starting our evaluation we performed a minimal functionality test and found that the latest version of WLM-Operator is no longer functional on recent operating systems. Despite our best efforts and reaching out to Sylabs, we were unable to reproduce the minimal examples in the repository. Therefore, WLM-Operator can be considered retired and we will not further consider it.

Kube-Slurm requires the installation of a Kubernetes cluster on all nodes as part of its deployment, which violates **RQ2** that it must be able to operate without root access. Therefore, we will not further consider Kube-Slurm.

This only leaves HPK and Bridge Operator as viable targets for further evaluation along with KSI, which is introduced in the next section. However, when testing Bridge Operator we ran into a number of issues, which we reported on Github and created a pull request [41] with our code adjustments.

B. Kind Slurm Integration (KSI) Design

Our main objectives of designing another approach for combining Slurm and Kubernetes were that it should follow the **Over** model and support all Kubernetes features. Following the **Over** model, KSI can be run strictly inside Slurm jobs without relying on external components. This was important to us, as our use case involved a multi-user HPC system in which the users of KSI would not be able to deploy a control plane outside of their Slurm jobs as it is required by Bridge Operator or HPK. Moreover, as we could not find any existing projects employing the **Over** model, we consider this a research gap.

We utilized rootless Kind [10] to create a script that receives a Kubernetes workload, initializes a cluster inside a Slurm job, executes the workload and then closes the cluster as the Slurm job ends. Before settling on rootless Kind, we also considered Minikube [42], K3D [43] and Usernetes [44] but found rootless Kind to be the most suitable.

Kind [45] was developed with local development and automatic testing of Kubernetes in mind. It can deploy a fully functional Kubernetes cluster on a single node by deploying a "node" image, which internally runs all containers belonging to the cluster as nested containers. From the perspective of the host system, only a single container is running for the control plane node of the cluster. Moreover, by deploying multiple "node" images on the same host, Kind can simulate a multi-node cluster.

For operating KSI inside of Slurm jobs without access to root permissions, which are required for regular container operation, we employ rootless Kind [10]. Rootless Kind relies on a container runtime that supports rootless container deployment, which typically relies on Cgroups v2 features in the Linux kernel. In KSI 1 [1] we relied on rootless Podman, which in turn relies on the shadow-utils package to provide subuids and subgids for user namespaces. Since then we have switched to rootless Containerd for KSI 2 as it has better support for rootless container networking. Nevertheless, the dependencies employed by Podman and Containerd to enable rootless mode, are similar if not identical. For example, a central component for both versions of KSI is RootlessKit [46], a fakeroot implementation specifically for rootless containers.

In order to deploy KSI, the nodes must provide a recent Linux operating system with support for Cgroups v2. Rootless Podman or rootless Containerd must be set up as well as slirp4netns [47], to provide networking for rootless containers. Slirp4netns is the default network driver for rootless Containerd and used to be the default for rootless Podman until Podman 5.0 [48] was released, which uses pasta [49].

However, the Podman documentation [50] states that employing slirp4netns can lead to degraded network performance. We confirmed this in our publication [1] and noted it as a significant shortcoming of KSI 1. In order to overcome this shortcoming we replaced slirp4netns with bypass4netns [15][16] for KSI 2.

Bypass4netns is an experimental project that employs `SECCOMP_IOCTL_NOTIF_ADDFD`, a Seccomp filter introduced into the Linux kernel with version 5.9. It builds on top of slirp4netns by capturing socket syscalls made by slirp4netns and executing them in the host network namespace. By doing so, the authors of bypass4netns effectively achieved the same network throughput for rootless containers as for rootful containers that set the `--net=host` flag. This enables the deployment of rootless containers with the same network performance as rootful containers without compromising on security by exposing the host network.

Bypass4netns can be used with rootless Docker, Podman and Nerdctl, which is the Containerd CLI. It has been integrated with rootless Containerd and Nerdctl such that it can be utilized together more easily compared to Podman and Docker. Due to this integration we have switched from rootless Podman to rootless Containerd as noted previously. It should also be

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

noted that Usernetes [44] also added support for bypass4netns.

When employing bypass4netns via Nerdctl, bypass4netns expects an annotation to be set via Nerdctl, which is checked by the bypass4netns daemon. When the daemon finds that annotation on a container, it enables accelerated networking via bypass4netns. In KSI 2, Kind is responsible for calling Nerdctl to deploy containers, however, Kind does not support passing through annotations. We therefore had to employ a workaround by wrapping the Nerdctl binary on the nodes with a script that injects the annotations.

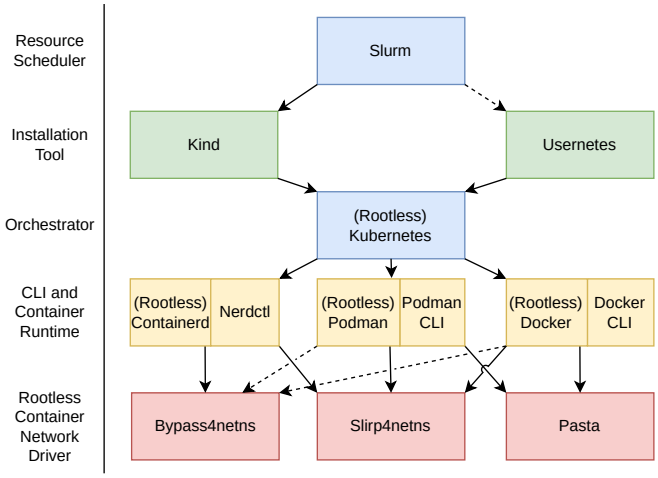


Figure 2. Overview of software options for setting up rootless Kubernetes on Slurm. From each row one option has to be used to assemble a viable software stack.

Figure 2 provides an overview of the components and alternatives in the software stack that can be employed for running rootless Kubernetes in Slurm jobs. Slurm is shown at the top as for the **Over** model the entire deployment is encapsulated in Slurm jobs. To deploy rootless Kubernetes, KSI employs Kind but we mentioned alternatives that also support rootless Kubernetes such as Usernetes, as shown in the figure, as well as K3D. In both cases, the goal of the tool is to deploy a rootless Kubernetes cluster.

The next layer in Figure 2 is the container runtime where the options are Containerd, Podman and Docker as each of them supports rootless containers. For Podman and Docker both the container runtime as well as the CLI tool share the same name, for Containerd, the associated CLI tool is called Nerdctl. KSI 1 utilized Podman but as discussed above, we switched to Containerd and Nerdctl for KSI 2. To complete the software stack a rootless container network driver is required, here the default in many cases is slirp4netns with bypass4netns being an experimental alternative.

Kind itself is not designed for multi-node clusters across multiple physical machines or VMs. Therefore, KSI 1 [1] was not able to achieve **RQ3**. However, in [1] we noted that a tool such as Kilo [51] or Liqo [17] could be used to enable mutli-node support. We have since then added Liqo to KSI 2 and enabled multi-node clusters such that **RQ3** can be satisfied.

Liqo operates by aggregating multiple Kubernetes clusters into a single cluster by representing each cluster as a virtual Kubelet in the main cluster. With this KSI 2 can be deployed across a number of nodes in a multi-node Slurm job and all the worker nodes each deploy a single node cluster. Then Liqo is installed and configured such that all single node clusters register themselves as virtual Kubelets on the main node. Liqo deploys a gateway pod in each cluster, which in our case means on each node that enables the clusters to automatically peer between each other. This allows for pod-to-pod communication between any of the nodes. Once the deployment and registration of all nodes through the main node has been completed, the main node is also no longer a single-point-of-failure for the cluster.

However, it should be noted that with this setup, every worker node runs a Kubernetes control plane instead of only Kubelet, as would be the case in a regular Kubernetes cluster. This adds overhead in terms of CPU and memory consumption on every node for running the Kubernetes control plane components.

The interface to run a Kubernetes workload via KSI is to submit it via Slurm, for example via `srun` in the following form:

```
srun -NX /bin/bash run-workload.sh
example-workload.sh /path/to/shared/folder
```

`X` in `-NX` specifies the number of nodes that should be requested from Slurm and used by the cluster. `run-workload.sh` is the main KSI script, which handles the provisioning of the rootless Kubernetes clusters via Kind and connecting them via Liqo. In order for Liqo to set up the multi-node cluster, it requires a shared folder across all nodes, for example an NFS share, which should be specified in place of `/path/to/shared/folder`. Finally, the actual workload should be specified in place of `example-workload.sh`. The workload script should use `kubectl` to create all components in Kubernetes that are needed and then wait for the workload to finish. Once the workload script returns, `run-workload.sh` will close the Kubernetes cluster and perform cleanup before stopping, which also ends the Slurm job.

The code for the KSI 1 implementation along with its documentation was released under the GPL-3.0 license on Github [25]. The code for KSI 2 with Liqo and bypass4netns integration was released under the same license on Github [9].

C. Performance Evaluation

To assess the performance overhead to answer **RQ4** we have broken down our benchmarking into the following factors:

- Startup time: Measured with a dummy workload
- CPU compute performance: Measured with Sysbench [52]
- Memory throughput: Measured with Stream [53]
- Storage throughput: Measured with Fio [54]
- Network latency: Measured with Netperf [55]
- Network bandwidth: Measured with iPerf3 [56]

We consider these as representative factors for user workloads that might be run through any of the tools under study.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

We performed two rounds of benchmarks. First, the comparative study of the combinations of Kubernetes and Slurm following the **Adjacent** model against our KSI 1 implementation. Second, the comparison of network drivers for the improved implementation of KSI 2. In both cases we measured bare metal baselines by running the benchmarks without Kubernetes.

The first set of benchmarks were run on two machines with hardware specifications as shown in Table I.

TABLE I. HARDWARE SPECIFICATIONS OF THE MACHINES IN THE FIRST SET OF BENCHMARKS.

CPU	Intel(R) Xeon(R) CPU E5-2695 v3
CPU Sockets	2
Cores per socket	14
Threads per core	2
Total threads	56
RAM	24 DIMMs DDR4 16 GB 1866 MHz
Total RAM	384.00 GB
Storage	1 Verbatim Vi550 S3 SATA Revision 3.2 SSD
Total storage	128.00 GB
Network interface	QLogic BRCM 10G/GbE 2+2P 57800-t rNDC

On these nodes we used software versions as shown in Table II. The Kubernetes version v1.27.3 was the most recent version at the time of the experiments and was used for the external cluster for the Bridge Operator as well as by KSI 1. HPK, however, is pinned to v1.25.0 in its code base. Furthermore, we disabled SELinux, swap and write caching to measure the respective factors more clearly.

TABLE II. SOFTWARE VERSIONS OF THE MACHINES IN THE FIRST SET OF BENCHMARKS.

Linux OS	CentOS Stream 9
Slurm	23.02.5
Podman	4.6.1
slirp4netns	1.2.2-1
Kind	0.20.0
Kubectl	v1.28.2
Kubernetes	v1.27.3
HPK Kubernetes	v1.25.0
shadow-utils	2:4.9-8

For the second benchmark to validate the improved network performance of bypass4netns for KSI 2 compared to other network drivers, we used two machines with hardware specifications as shown in Table III.

TABLE III. HARDWARE SPECIFICATIONS OF THE MACHINES IN THE SECOND SET OF BENCHMARKS.

CPU	Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz
CPU Sockets	2
Cores per socket	20
Threads per core	2
Total threads	80
RAM	DIMM DDR4 Synchronous 2666 MHz
Total RAM	192 GB
Storage	Served from RAM due to Warewulf
Total storage	94 GB
Network interface	Intel 82599ES 10-Gigabit SFI/SFP+

On these we used software versions as listed in Table IV. Most significantly we switched to Rocky Linux from CentOS, replaced Podman with Nerdctl and Containerd and added bypass4netns. Also we used newer versions of Kind and Kubernetes.

TABLE IV. SOFTWARE VERSIONS OF THE MACHINES IN THE SECOND SET OF BENCHMARKS.

Linux OS	Rocky Linux 9 (Kernel 5.14)
Slurm	23.02.5
Nerdctl	v2.1.3
Containerd	v1.7.27
RootlessKit	v2.3.5
slirp4netns	v1.3.2
bypass4netns	v0.4.2
Kind	0.29.0
Kubectl	Client Version: v1.33.2
Kubernetes	Server Version: v1.33.2
Libseccomp	2.5.2
Liqoctl	v1.0.1

D. Project State Evaluation

Evaluating the maintainability and usability of software has been studied extensively [57][58] with many tools and methods having been proposed. For this work, in order to answer **RQ5** and **RQ6**, we have to consider what methods to employ.

In order to grade usability we consider the state of the available documentation as well as the difficulty of setting up and operating the respective tools for an assumed non-expert user based on our own experience of working with the tools during this study. For grading maintainability we reviewed the state of the code repositories based on their complexity, whether they have been kept up-to-date and how well issues had been addressed. Moreover, we consider that a code base that does a comparatively simple job while relying on more well maintained dependencies is itself more maintainable than a larger code that has more moving parts that may require maintenance.

We acknowledge that more sophisticated methods are available but consider our approach sufficient to compare the three projects under study on a three point grading schema.

IV. RESULTS

The commit hashes of the implementation versions used in our tests are as follows:

- Bridge Operator: `56334fa57caf2de28df6ff76df8a6e6232021421`
- HPK: `a902acbf2436e8a85a4620fddfa5745523f443d4`
- KSI 1: `780ef3a0562ad4bb12611f9ef43fa743fe0277d0` [25]
- KSI 2: `ea0457a97d056d32947d6e99538aac6e174e9213` [9]

A. Functional Requirements

For Bridge Operator and HPK the answer to **RQ1**, **RQ2** and **RQ3** is yes. For KSI 1 as well except for **RQ3**. This is solved in KSI 2 via the integration with Liqo.

B. Project State

a) Bridge Operator: When submitting a workload through Bridge Operator, it requires the user to create an instance of the CRD BridgeJob. With that no understanding of Slurm by the user is required. However, the available documentation for Bridge Operator is limited with some examples not working

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

such that a patch was required to make it work [41]. While the project depends only on the Slurm REST API, giving it a stable foundation, the project itself seems abandoned with no activity after late 2022. Due to this, we rate it ○ in both usability and maintainability. We would have rated + for usability if the examples were all functional and for maintainability if the project was actively being maintained.

b) HPK: Similar to Bridge Operator, HPK can be controlled directly through `kubectl` without additional understanding of Slurm by the user. Nevertheless, while its documentation is also limited, after we had completed our experiments [35] was released along with `v0.1.2` of HPK containing a number of bug fixes. This shows that the project is being actively developed, moreover, when we ran into issues, we quickly received community support from the maintainers. With this we rate the maintainability as + and the usability as ○ because in addition to the points mentioned above, HPK does not support certain Kubernetes features, most notably `kubectl exec` and services, such that users need to work around these limitations.

c) KSI: Unlike the other two tools, KSI is started via Slurm as it has no active component outside of the Slurm job. Its usage is documented with several examples and it depends on Kind and either Podman for the original implementation and Containerd and Nerdctl for the improved version. Kind, Podman, Containerd and Nerdctl are well maintained projects. However, KSI 2 also depends on experimental features of Nerdctl as well as bypass4netns, which is also not in a stable state yet. Nevertheless, KSI delivers a feature complete Kubernetes cluster via a set of scripts making it both usable and maintainable so we rate KSI + for both factors. However, as KSI is our own creation, we cannot claim that this evaluation is unbiased and should be regarded as such.

C. Performance

The benchmarking scripts, as well as the raw test data, are available on Github [59]. Each benchmark was repeated 10 times to minimize random error with the standard deviation shown in the graphs. All benchmarks, except for those covered in Section IV-C6, were done as part of the first set of benchmarks with the implementation of KSI 1. Section IV-C6 covers the investigation of the network performance of KSI 2 with bypass4netns compared to slirp4netns, which KSI 1 used, and forms the second set of benchmarks.

TABLE V. WORKLOAD STARTUP TIME, LOWER IS BETTER.

Integration Approach	Startup Time in s
Bare Metal	0.141
Bridge Operator	2.725
HPK	2.497
KSI 1	53.921

1) Startup Time: The startup delays given in Table V have negligible standard deviation and show that Bridge Operator and HPK start a workload in 2 to 3 seconds while KSI 1 requires almost one minute. This result is as expected since Bridge Operator and HPK already have an active Kubernetes cluster running before they submit their Slurm job while KSI has to set up a Kubernetes cluster from scratch. Considering that HPC workloads often run for multiple hours, one minute extra start up time is not great but acceptable. Due to this we rate Bridge Operator and HPK with + and KSI with ○.

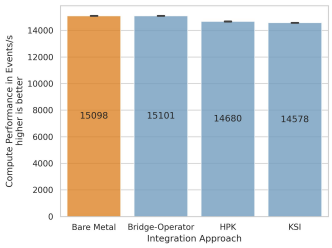


Figure 3. CPU compute performance results using Sysbench. Data was collected with KSI 1.

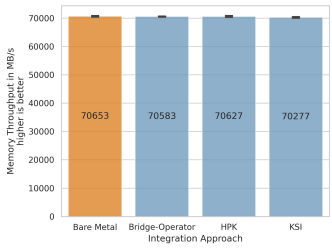


Figure 4. Memory throughput results using Stream. Data was collected with KSI 1.

2) Compute Performance: Figure 3 shows that Bridge Operator and bare metal are effectively on the same performance level while HPK and KSI 1 are slightly lower than bare metal (2.7% and 3.4%, respectively). The difference arises due to the virtualization overhead and additional active components running for HPK and KSI but is overall negligible so we rate all approaches with +.

3) Memory Performance: Figure 4 shows similar to Figure 3 only minor differences for HPK and KSI 1 due to the additional active components and virtualization such that we also rate all with + here.

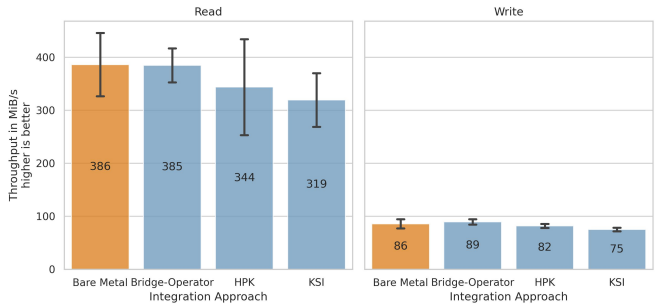


Figure 5. Storage throughput results using Fio and sequential operations. Data was collected with KSI 1.

4) Storage Performance: The sequential read and write shown in Figure 5 shows a similar pattern as the random read and write shown in Figure 6 with Bridge Operator being on the same level as bare metal and HPK and KSI 1 lacking behind. More specifically HPK is about 11% slower in sequential and 5% slower in random reading and KSI 1 is overall 17% slower in reading and 13% slower in writing than bare metal. These differences can also be attributed to the additional virtualization and overall resource consumption. While 17% slower reading is not good we rate it still as acceptable so HPK and KSI are rated as ○ and Bridge Operator as +.

5) Network Performance: For these benchmarks, the respective tool executed a workload containing a test client that

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

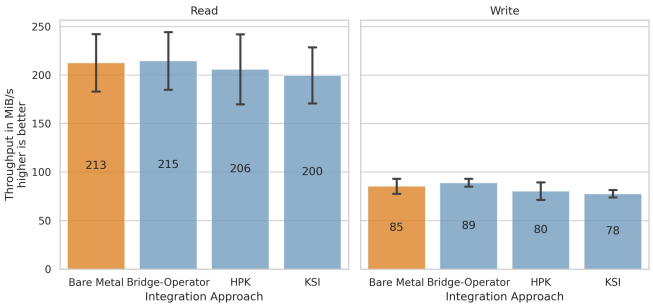


Figure 6. Storage throughput results using Fio and randomized operations. Data was collected with KSI 1.

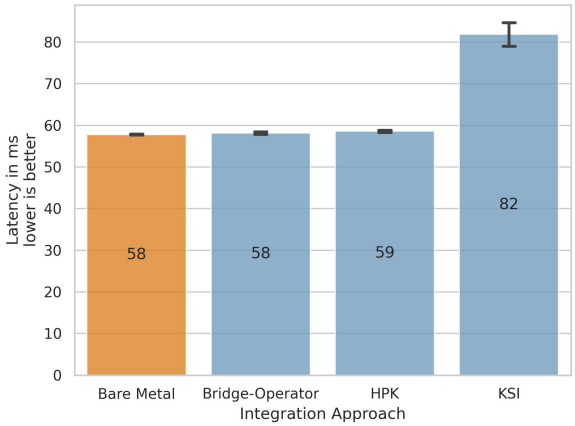


Figure 7. Network latency results using Netperf for the Kubernetes-Slurm integration solutions under study. Data was collected with KSI 1, which used slirp4netns.

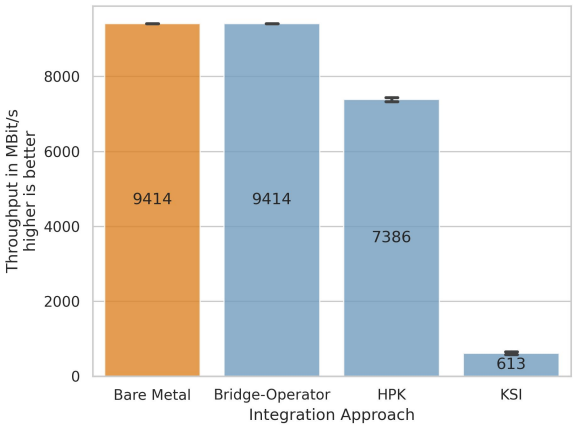


Figure 8. Network bandwidth results using iPerf3 for the Kubernetes-Slurm integration solutions under study. Data was collected with KSI 1, which used slirp4netns.

executed the network benchmark against a server running on the other of the two nodes in our test setup. What is shown as the bare metal latency and throughput are therefore the latency and peak throughput between the two nodes. Figure 7 shows network latency with all solutions on the same level except for KSI 1, which is 42% slower than bare metal. In Figure 8 the network throughput is even worse for KSI 1 with HPK already being 21% slower than bare metal, KSI 1 is 93.5% slower. As KSI 1 operates via rootlesss Podman, it uses slirp4netns as its driver, which according to the Podman documentation [50] results in degraded performance compared to rootful Podman networking.

Since our experiments concluded, pasta had replaced slirp4netns as the default network driver for rootless Podman, which promises better performance but initial tests could not show a significant overall improvement [60]. Our ratings are + for Bridge Operator, ○ for HPK and − for KSI 1.

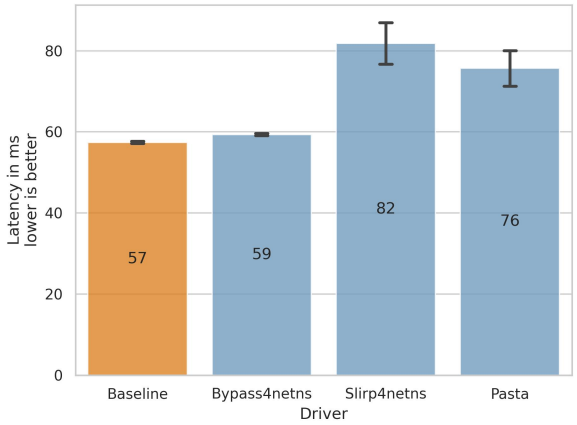


Figure 9. Network latency results using Netperf for KSI 2 using various rootless container network drivers. All variants were tested with Nerdctl except for pasta, which was tested with Podman 5.x.

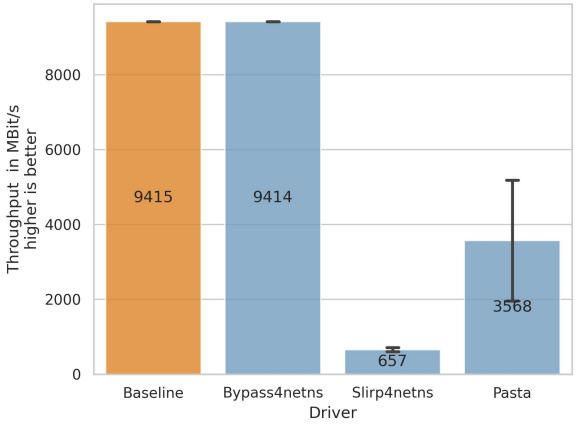


Figure 10. Network bandwidth results using iPerf3 for KSI 2 using various rootless container network drivers. All variants were tested with Nerdctl except for Pasta, which was tested with Podman 5.x.

6) Network Performance with Bypass4netns: In order to evaluate whether adding bypass4netns to KSI 2 improved the network performance, we compared its performance to slirp4netns and no containerization as a baseline. For all of these we used Nerdctl with rootless Containerd. Moreover, we also investigated pasta with Podman as pasta had replaced

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

slirp4netns in Podman 5.0 in order to improve network performance of rootless containers. These measurements were conducted in the second set of benchmarks utilizing the hardware and software as listed in Tables III and IV.

The network latency shown in Figure 9 closely mirrors the results from the first round of benchmarks in Figure 7. Most notably, bypass4netns achieves effectively identical latency to the baseline. Pasta on the other hand shows slight improvements over slirp4netns but does not provide a silver bullet solution.

For network throughput Figure 10 also closely mirrors Figure 8 from the first round of benchmarks. Bypass4netns has also in this area caught up to the baseline, which shows that by employing bypass4netns, the network limitations of the KSI 1 implementation can be overcome.

However, as the hardware used in our experiments was limited to 10 Gbps, it is possible that bypass4netns would not be able to keep up with the baseline on more powerful network hardware. Nevertheless, based on the benchmarks done by the authors of bypass4netns [16], we can estimate that bypass4netns will be close to the baseline within a few percentage points even when using more powerful hardware. Our rating for the network performance of KSI 2 is +.

D. Evaluation

TABLE VI. PROJECT ASSESSMENT REGARDING QUALITY REQUIREMENTS. *KSI 1* REPRESENTS THE ORIGINAL KSI IMPLEMENTATION AND *KSI 2* THE IMPROVED KSI IMPLEMENTATION WITH LIQO AND BYPASS4NETNS. * SELF-EVALUATION OF KSI IS NOT UNBIASED.

Project	**RQ4** Startup	**RQ4** Comp.	**RQ4** Storage	**RQ4** Net.	**RQ5** Usab.	**RQ6** Maintainab.
B-O	+	+	+	+	○	○
HPK	+	+	○	○	○	+
KSI 1	○	+	○	−	+*	+*
KSI 2	○	+	○	+	+*	+*

Table VI summarizes the ratings we have assigned throughout this section with Startup, Compute, Storage and Network performance all aiming at **RQ4** and Usability and Maintainability aiming at **RQ5** and **RQ6**, respectively.

Bridge Operator has shown performance close or identical to bare metal, which is as expected since it effectively submits a Slurm job through Kubernetes and does not start additional software in that Slurm job. This brings some limitations as it is not actually running a given workload using Kubernetes. Nevertheless, it has presented itself as a valid approach for extending a Kubernetes cluster via access to a Slurm cluster.

HPK provides a good middle ground for running Kubernetes jobs on Slurm with some performance deficiencies compared to bare metal. If WLM-Operator would be functional, we would have probably seen similar performance to HPK as WLM-Operator is based on Singularity and HPK is based on Apptainer and both projects still share the majority of their implementation. While HPK does not support all Kubernetes features, e.g., services and `kubectl exec` are not supported, it provides a solid choice for natively running Kubernetes workloads through Slurm.

KSI is functionally the most complete Kubernetes environment within a Slurm job and requires no external parts to be started and kept running outside of it. For the implementation of KSI 1 this comes with performance costs, as KSI 1 shows the weakest performance in all benchmarks, especially in startup time and networking. The slow startup time is understandable as KSI 1 has to bootstrap the Kubernetes control plane and cannot rely on an existing Kubernetes cluster. For network performance, KSI 1 relies on slirp4netns, which is known for causing performance degradation [50].

The KSI 2 implementation utilizes bypass4netns, which overcame the limited network performance of the original implementation. This puts the KSI 2 implementation on the same level as Bridge Operator as both were effectively on the same level as the baseline or bare metal measurements.

Nevertheless, in order for our KSI 2 implementation to achieve these performance levels, bypass4netns must be set up on each node. Bypass4netns requires access to features only available on Linux kernel version 5.9 and newer, which might not be available in all HPC environments. For instance, only RHEL-like 9 and newer operating systems, such as Rocky Linux 9, include a recent enough kernel version for bypass4netns. Moreover, security consideration as mentioned by the authors of bypass4netns [16] apply.

All the projects suffer from being either only proof-of-concept implementations, not being maintained or not being properly documented such that none of them provide a production ready solution for running Kubernetes inside of Slurm jobs.

V. CONCLUSION AND FUTURE WORK

In this paper, we analyzed the state of solutions for combining Slurm and Kubernetes with the goal to enable dynamic computation between either environment. We focused on a subset of the available solutions to support our use case of running Kubernetes workloads on an existing Slurm cluster.

For this purpose, we improved upon our own solution KSI 1, which was originally based on Kind and rootless Podman and was able to deploy a fully functional Kubernetes cluster inside a Slurm job. We improved KSI 1 by switching from rootless Podman to rootless Nerdctl and employing bypass4netns for container networking instead of Podman's default network driver slirp4netns. Moreover, we added Liqo to enable multi-node clusters for KSI 2.

From the available solutions we took a closer look at Bridge Operator, HPK and our own solution, KSI, and found that they fulfill our functional requirements, except for KSI 1, for which multi-node support had not been implemented yet. KSI 2, our improved implementation, supports multi-node clusters. We further evaluated the performance of each solution and reviewed the state of their respective implementations.

We found that Bridge Operator delivers effectively bare metal performance equal to directly running a job through Slurm as this is effectively what Bridge Operator does. HPK established itself as a middle ground solution, providing an almost fully functional Kubernetes cluster inside a Slurm job

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

with minor performance overhead. Our solution, KSI 1, showed slightly higher overhead compared to HPK and significantly less network throughput in its original implementation. Via the improvements presented in this work, we were able to achieve the same level of network performance as Bridge Operator. KSI 2 provides the most feature complete Kubernetes clusters compared to the other solutions. However, to achieve its fast network throughput it relies on recent Linux kernel features that might not be available in every HPC center. Moreover, it requires carefully preparing the system environments on every node that should run KSI 2 by setting up the respective software stack.

We conclude that by overcoming the two significant shortcomings of KSI 1, it is now able to deliver both feature complete Kubernetes environments as well as excellent performance. The next steps for KSI 2 include improving its overall usability and stability as well as comparing it to other solutions such as Usernetes.

Acknowledgments

This work was supported by the Federal Ministry of Education and Research (BMBF), Germany under the AI service center KISSKI (grant no. 01IS22093A and 01IS22093B). Moreover, additional work was funded by the European Union under the Horizon Europe initiative via grant no. 101092582 for the DECICE project.

References

[1] J. Decker, S. Metje, and J. Kunkel, "Running Kubernetes Workloads on Rootless HPC Systems using Slurm", presented at the CLOUD COMPUTING 2025, The Sixteenth International Conference on Cloud Computing, GRIDs, and Virtualization, Apr. 6, 2025, pp. 100–107, ISBN: 978-1-68558-258-6.

[2] "CNCF Annual Survey 2023", CNCF, Apr. 9, 2024, [Online]. Available: https://www.cncf.io/reports/cncf-annual-survey-2023/ (visited on 2024.12.30).

[3] "9 Insights on Real-World Container Use | Datadog", [Online]. Available: https://web.archive.org/web/20230318234844/https://www.datadoghq.com/container-report/ (visited on 2024.12.30).

[4] A. B. Yoo, M. A. Jette, and M. Grondona, "SLURM: Simple Linux Utility for Resource Management", in *Job Scheduling Strategies for Parallel Processing*, D. Feitelson, L. Rudolph, and U. Schwiegelshohn, Eds., Berlin, Heidelberg: Springer, 2003, pp. 44–60, ISBN: 978-3-540-39727-4. DOI: 10.1007/10968987_3.

[5] "Volcano-sh/volcano", Volcano, Dec. 30, 2024, [Online]. Available: https://github.com/volcano-sh/volcano (visited on 2024.12.30).

[6] "Scaling Kubernetes to 7,500 Nodes", Jan. 2021, [Online]. Available: https://openai.com/blog/scaling-kubernetes-to-7500-nodes/ (visited on 2021.02.12).

[7] A. M. Beltre, P. Saha, M. Govindaraju, A. Younge, and R. E. Grant, "Enabling hpc workloads on cloud infrastructure using kubernetes container orchestration mechanisms", in *2019 IEEE/ACM International Workshop on Containers and New Orchestration Paradigms for Isolated Environments in HPC (CANOPIE-HPC)*, Nov. 2019, pp. 11–20. DOI: 10.1109/CANOPIE-HPC49598.2019.00007.

[8] T. Wickberg, "Slurm and/or/vs Kubernetes", Schedmd, [Online]. Available: https://slurm.schedmd.com/SC23/Slurm-and-or-vs-Kubernetes.pdf (visited on 2025.08.06).

[9] "Gwdg/pub-2025-ksi", Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Aug. 1, 2025, [Online]. Available: https://github.com/gwdg/pub-2025-ksi (visited on 2025.08.01).

[10] "Kind – Rootless", [Online]. Available: https://kind.sigs.k8s.io/docs/user/rootless/ (visited on 2024.12.30).

[11] B. Lublinsky, E. Jennings, and V. Spišaková, "A kubernetes 'bridge' operator between cloud and external resources", in *2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 2023, pp. 263–269. DOI: 10.1109/ICCCBDA56900.2023.10154770.

[12] Staff, "Introducing HPC Affinities to the Enterprise: A New Open Source Project Integrates Singularity and Slurm via Kubernetes", Sylabs, May 7, 2019, [Online]. Available: https://sylabs.io/2019/05/introducing-hpc-affinities-to-the-enterprise-a-new-open-source-project-integrates-singularity-and-slurm-via-kubernetes/ (visited on 2024.12.30).

[13] K. Peterson, "Kalenpeterson/kube-slurm", Aug. 17, 2024, [Online]. Available: https://github.com/kalenpeterson/kube-slurm (visited on 2024.12.30).

[14] A. Chazapis, F. Nikolaidis, M. Marazakis, and A. Bilas, "Running kubernetes workloads on HPC", in *High Performance Computing*, A. Bienz, M. Weiland, M. Baboulin, and C. Kruse, Eds., ser. Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2023, pp. 181–192, ISBN: 978-3-031-40843-4. DOI: 10.1007/978-3-031-40843-4_14.

[15] N. Matsumoto and A. Suda, "Accelerating TCP/IP Communications in Rootless Containers by Socket Switching", 2022.

[16] N. Matsumoto and A. Suda, "Bypass4netns: Accelerating TCP/IP Communications in Rootless Containers", Feb. 1, 2024. DOI: 10.48550/arXiv.2402.00365. arXiv: 2402.00365 [cs], pre-published.

[17] "Liqotech/liqo", LiqoTech, Jul. 22, 2025, [Online]. Available: https://github.com/liqotech/liqo (visited on 2025.07.24).

[18] S. Metje, *Running Kubernetes Workloads on Rootless HPC Systems using Slurm*, GRO.data, Jan. 9, 2024. DOI: 10.25625/GDFCFP.

[19] F. Liu, K. Keahey, P. Riteau, and J. Weissman, "Dynamically negotiating capacity between on-demand and batch clusters", in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, ser. SC '18, Dallas, Texas: IEEE Press, Nov. 11, 2018, pp. 1–11.

[20] B. Wu, M. Hu, S. Qin, and J. Jiang, "Research on fusion scheduling based on Slurm and Kubernetes", in *International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2024)*, vol. 13403, SPIE, Nov. 18, 2024, pp. 476–485. DOI: 10.1117/12.3051639.

[21] G. Zervas, A. Chazapis, Y. Sfakianakis, C. Kozanitis, and A. Bilas, "Virtual clusters: Isolated, containerized HPC environments in kubernetes", in *High Performance Computing. ISC High Performance 2022 International Workshops*, H. Anzt, A. Bienz, P. Luszczek, and M. Baboulin, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2022, pp. 347–357, ISBN: 978-3-031-23220-6. DOI: 10.1007/978-3-031-23220-6_24.

[22] T. Menouer, N. Greneche, C. Cérin, and P. Darmon, "Towards an Optimized Containerization of HPC Job Schedulers Based on Namespaces", in *Network and Parallel Computing*, C. Cérin, D. Qian, J.-L. Gaudiot, G. Tan, and S. Zuckerman, Eds., Cham: Springer International Publishing, 2022, pp. 144–156, ISBN: 978-3-030-93571-9. DOI: 10.1007/978-3-030-93571-9_12.

[23] C. Cérin, N. Greneche, and T. Menouer, "Towards Pervasive Containerization of HPC Job Schedulers", in *2020 IEEE 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, Sep. 2020, pp. 281–288. DOI: 10.1109/SBAC-PAD49847.2020.00046.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[24] "SlinkyProject/slurm-operator", SlinkyProject, Dec. 26, 2024, [Online]. Available: https://github.com/SlinkyProject/slurm-operator (visited on 2024.12.30).

[25] S. Metje, "Kubernetes Slurm Integration based on Kind", 2023, [Online]. Available: https://github.com/soerenmetje/kind-slurm-integration.

[26] P. Liu and J. Guitart, *Fine-grained scheduling for containerized HPC workloads in kubernetes clusters*, Nov. 21, 2022. DOI: 10.48550/arXiv.2211.11487. arXiv: 2211.11487[cs].

[27] D. Medeiros, J. Wahlgren, G. Schieffer, and I. Peng, "Kub: Enabling elastic HPC workloads on containerized environments", in *2023 IEEE 35th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, ISSN: 2643-3001, Oct. 2023, pp. 219–229. DOI: 10.1109/SBAC-PAD59825.2023.00031.

[28] "PRIMAGE / hpc-connector · GitLab", GitLab, Feb. 22, 2023, [Online]. Available: https://gitlab.com/primageproject/hpc-connector (visited on 2025.01.02).

[29] N. Zhou *et al.*, "Container orchestration on HPC systems through Kubernetes", *Journal of Cloud Computing*, vol. 10, no. 1, p. 16, Feb. 22, 2021, ISSN: 2192-113X. DOI: 10.1186/s13677-021-00231-z.

[30] G. Staples, "TORQUE resource manager", in *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, ser. SC '06, New York, NY, USA: Association for Computing Machinery, Nov. 11, 2006, 8–es, ISBN: 978-0-7695-2700-0. DOI: 10.1145/1188455.1188464.

[31] "Sylabs/wlm-operator", Sylabs Inc., Nov. 5, 2024, [Online]. Available: https://github.com/sylabs/wlm-operator (visited on 2025.01.02).

[32] "Sylabs/singularity-cri", Sylabs Inc., Mar. 1, 2024, [Online]. Available: https://github.com/sylabs/singularity-cri (visited on 2025.01.02).

[33] B. Lublinsky, E. Jennings, and V. Spišaková, "A Kubernetes 'Bridge' operator between cloud and external resources", Jul. 6, 2022. DOI: 10.48550/arXiv.2207.02531. arXiv: 2207.02531 [cs], pre-published.

[34] "Bridge-Operator/kubeflow at main · IBM/Bridge-Operator", [Online]. Available: https://github.com/IBM/Bridge-Operator/tree/main/kubeflow (visited on 2025.01.02).

[35] "CARV-ICS-FORTH/HPK", Computer Architecture and VLSI Systems (CARV) Laboratory, Dec. 26, 2024, [Online]. Available: https://github.com/CARV-ICS-FORTH/HPK (visited on 2025.01.02).

[36] A. Chazapis, E. Maliaroudakis, F. Nikolaidis, M. Marazakis, and A. Bilas, "Running Cloud-native Workloads on HPC with High-Performance Kubernetes", Sep. 25, 2024. DOI: 10.48550/arXiv.2409.16919. arXiv: 2409.16919 [cs], pre-published.

[37] "Apptainer/apptainer", The Apptainer Container Project, Dec. 30, 2024, [Online]. Available: https://github.com/apptainer/apptainer (visited on 2025.01.02).

[38] "Virtual-kubelet/virtual-kubelet", virtual kubelet, Feb. 24, 2025, [Online]. Available: https://github.com/virtual-kubelet/virtual-kubelet (visited on 2025.02.24).

[39] "Flannel-io/flannel", flannel-io, Feb. 25, 2025, [Online]. Available: https://github.com/flannel-io/flannel (visited on 2025.02.25).

[40] "Flannel-io/cni-plugin", flannel-io, Jan. 31, 2025, [Online]. Available: https://github.com/flannel-io/cni-plugin (visited on 2025.02.25).

[41] "Fix #2 #3 #6 by soerenmetje · Pull Request #4 · IBM/Bridge-Operator", [Online]. Available: https://github.com/IBM/Bridge-Operator/pull/4 (visited on 2025.01.02).

[42] "Kubernetes/minikube", Kubernetes, Jan. 2, 2025, [Online]. Available: https://github.com/kubernetes/minikube (visited on 2025.01.02).

[43] "K3d-io/k3d", k3d, Jan. 2, 2025, [Online]. Available: https://github.com/k3d-io/k3d (visited on 2025.01.02).

[44] "Rootless-containers/usernetes", rootless-containers, Dec. 30, 2024, [Online]. Available: https://github.com/rootless-containers/usernetes (visited on 2025.01.02).

[45] "Kind", [Online]. Available: https://kind.sigs.k8s.io/ (visited on 2025.02.21).

[46] "Rootless-containers/rootlesskit", rootless-containers, Jul. 25, 2025, [Online]. Available: https://github.com/rootless-containers/rootlesskit (visited on 2025.07.25).

[47] "Rootless-containers/slirp4netns", rootless-containers, Feb. 23, 2025, [Online]. Available: https://github.com/rootless-containers/slirp4netns (visited on 2025.02.25).

[48] "Releases · containers/podman", GitHub, [Online]. Available: https://github.com/containers/podman/releases (visited on 2025.02.25).

[49] "Passt - Plug A Simple Socket Transport", [Online]. Available: https://passt.top/passt/about/ (visited on 2025.02.25).

[50] "Podman/docs/tutorials/rootless_tutorial.md at main · containers/podman", GitHub, [Online]. Available: https://github.com/containers/podman/blob/main/docs/tutorials/rootless_tutorial.md (visited on 2025.02.21).

[51] L. S. Marín, "Squat/kilo", Dec. 24, 2024, [Online]. Available: https://github.com/squat/kilo (visited on 2025.01.02).

[52] A. Kopytov, "Akopytov/sysbench", Jan. 2, 2025, [Online]. Available: https://github.com/akopytov/sysbench (visited on 2025.01.02).

[53] J. Hammond, "Jeffhammond/STREAM", Dec. 24, 2024, [Online]. Available: https://github.com/jeffhammond/STREAM (visited on 2025.01.02).

[54] J. Axboe, "Flexible I/O Tester", 2022, [Online]. Available: https://github.com/axboe/fio (visited on 2025.01.02).

[55] "HewlettPackard/netperf", Hewlett Packard Enterprise, Dec. 10, 2024, [Online]. Available: https://github.com/HewlettPackard/netperf (visited on 2025.01.02).

[56] "Esnet/iperf", ESnet: Energy Sciences Network, Jan. 2, 2025, [Online]. Available: https://github.com/esnet/iperf (visited on 2025.01.02).

[57] L. Ardito, R. Coppola, L. Barbato, and D. Verga, "A Tool-Based Perspective on Software Code Maintainability Metrics: A Systematic Literature Review", *Scientific Programming*, vol. 2020, no. 1, p. 8 840 389, 2020, ISSN: 1875-919X. DOI: 10.1155/2020/8840389.

[58] K. A. Dawood *et al.*, "Towards a unified criteria model for usability evaluation in the context of open source software based on a fuzzy Delphi method", *Information and Software Technology*, vol. 130, p. 106 453, Feb. 1, 2021, ISSN: 0950-5849. DOI: 10.1016/j.infsof.2020.106453.

[59] "Gwdg/pub-2025-ksi-evaluation", Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Aug. 1, 2025, [Online]. Available: https://github.com/gwdg/pub-2025-ksi-evaluation (visited on 2025.08.01).

[60] "Rootless network performance (pasta vs slirp4netns) · containers/podman · Discussion #22559", GitHub, [Online]. Available: https://github.com/containers/podman/discussions/22559 (visited on 2025.01.03).

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Designing Inclusive Immersive Experiences: Intergenerational Codesign of Heritage and Underwater Extended Reality for Older Adults

Marius N. Varga
School of Nursing and Midwifery
University of Plymouth
Plymouth, United Kingdom
marius.varga@plymouth.ac.uk

Oksana Hagen
School of Engineering,
Computing and Mathematics
University of Plymouth
Plymouth, United Kingdom
oksana.hagen@plymouth.ac.uk

Rory Baxter
School of Nursing and Midwifery
University of Plymouth
Plymouth, United Kingdom
rory.baxter@plymouth.ac.uk

Alejandro Veliz Reyes
School of Art, Design and Architecture
University of Plymouth
Plymouth, United Kingdom
alejandro.velizreyes@plymouth.ac.uk

Ray B Jones
School of Nursing and Midwifery
University of Plymouth
Plymouth, United Kingdom
ray.jones@plymouth.ac.uk

Amir Aly
School of Engineering,
Computing and Mathematics
University of Plymouth
Plymouth, United Kingdom
amir.aly@plymouth.ac.uk

Dena Bazazian
School of Engineering,
Computing and Mathematics
University of Plymouth
Plymouth, United Kingdom
dena.bazazian@plymouth.ac.uk

Swen Gaudl
Department of Applied IT
University of Gothenburg
Gothenburg, Sweden
swen.gaudl@ait.gu.se

***Abstract*—Despite rapid advances in immersive Extended Reality technologies, creating inclusive and accessible experiences for older adults remains challenging. Many Extended Reality applications fail to account for the physical, cognitive, and ergonomic needs of older users, resulting in low engagement and digital exclusion. Cultural heritage and underwater environments offer rich contexts for meaningful interaction and wellbeing. This study explored how intergenerational codesign can inform the development of immersive Extended Reality applications for older adults. Two case studies were undertaken within the ICONIC project: a Heritage Extended Reality experience recreating Cotehele, a National Trust heritage site, and an Underwater Telepresence experience simulating marine environments. The objective was to identify design principles supporting accessibility, comfort, and engagement. Sixteen codesign workshops were conducted with 24 older and 12 younger adults across five iterative stages: framing, ideation, prototyping, digital design, and testing. Mixed-methods analysis combined thematic coding of qualitative feedback with non-parametric statistical tests to assess design priorities. Participants prioritised immersion, interactivity, and accessibility. Simplified locomotion and one-button control schemes enhanced usability, while ergonomic modifications to headsets and controllers improved comfort. For Underwater Telepresence, dual "relaxation" and "learning" modes broadened appeal. Social features were consistently ranked lower than immersive and educational elements. Intergenerational codesign offers an effective approach to inclusive Extended Reality development and these insights inform future applications addressing digital exclusion and promoting equitable access to cultural and environmental experiences. In this paper, we showed that simplified controls, ergonomic design, and affordable hardware improve engagement and confidence among older users for immersive heritage and underwater telepresence experiences.**

Keywords- extended reality; intergenerational codesign; heritage; underwater; digital exclusion; immersive; telepresence.

I. Introduction

Building on our previous work on immersive cultural heritage experiences [1] presented at the The Second International Conference on Artificial Intelligence and Immersive Virtual Reality (AIVR2025) [2], this paper extends the investigation of how Virtual Reality (VR) can support meaningful engagement across diverse audiences with a special focus on the older adults. In our earlier study, we explored how age-related differences influenced user preferences and modes of interaction within virtual environments, and we discussed how the codesign process shaped the overall experience for both heritage sites and telepresence for underwater environments. The current work expands these findings by presenting quantitative and qualitative evidence collected during the evaluation stage of the codesign process and how this evidence influences future development.

Advances in immersive technology such as Extended Reality (XR) have opened the door to new opportunities creating novel ways of exploring abstract concepts such as the visualisation of a Tadpole spinal cord firing mechanism [3] or representing complex numbers as shapes [4] but also accessing an engaging with museums and heritage sites [5] through recreating real world locations in a Virtual environment. Evidence suggests that younger audiences prefer Virtual Reality (VR) as a learning environment for cultural heritage [6]. This

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

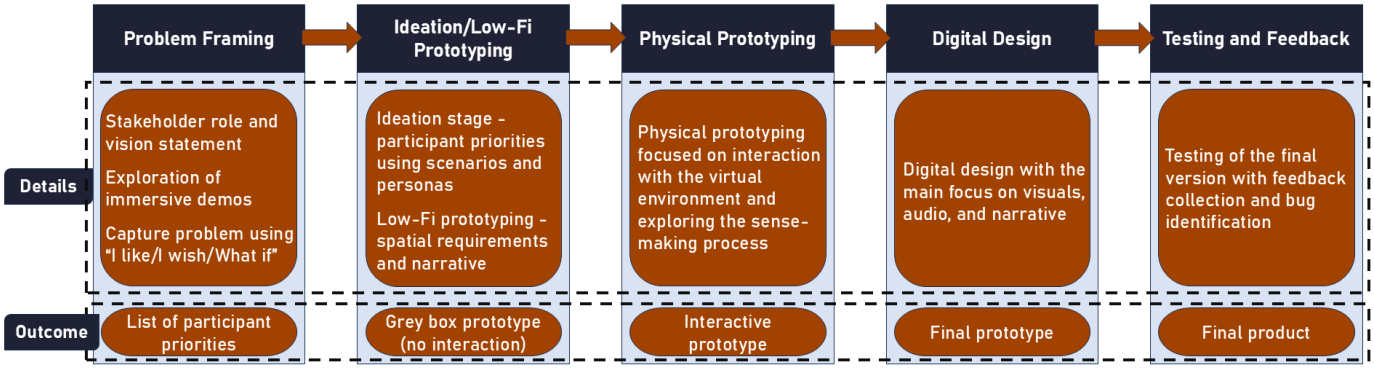


Figure 1. Timeline for the HXR and UT work package. The timeline is split in five distinct stages with an approximate one month duration for each stage. This approach is common to all ICONIC's work packages.

does not imply technological exclusion for older audiences, it highlights the need to go beyond usability testing and consider their interests and requirements for newly created content [7]. Although younger audiences are potentially drawn to the novelty and interactivity of VR, older audiences might engage more deeply with content that resonates with their life experiences and memories.

Through the Intergenerational Codesign of Novel Technologies In Coastal Communities (ICONIC) project, our aim was to give codesign participants (contributors), both young and old, a voice in the creation process and help them integrate their needs, suggestions and requirements into the design of novel technologies. The general aim of the ICONIC project was to develop four novel technologies through intergenerational codesign that would help connect digitally disadvantaged older people to local heritage and the environment. From the expertise of our team and our coastal context, we had nominated four general areas of technology: extended reality for heritage, telepresence for underwater, social games, and voice-AI over the telephone. This paper will focus on the intergenerational codesign approach to develop a Heritage Extended Reality (HXR) application and an Underwater Telepresence (UT) application.

Ijaz et al. [8] highlighted ten categories of design considerations for older adults that focused on users and physical configuration, hardware usage and the design of the immersive application. Through our codesign process we had the opportunity to address the majority of the categories such as: *onboarding and assistance* through supplementary sessions in order to familiarise the contributors with the hardware and the concept of VR; *safety* with support for contributors to explore VR standing or sitting in the presence of a researcher; *visuals* designed and created to capture the sense of being the physical heritage site; *audio* with the implementation of spatial audio; *personalisation* customisation of the VR headset and the controllers to help users with reduced mobility; *usability* with custom interaction metaphors for engaging with the virtual environment; *engagement* adding a gamified experience through interaction with historical artefacts; *minimise side effects* through support and clear instructions especially when testing new control mappings or unique locomotion techniques. *Embodiment* was addressed through the implementation of localised walking complemented by teleportation. In addition, it also includes automatic adjustment of the user's height when they put on the VR headset although we did not use an avatar to represent the user's body. *Realism* was addressed using actual measurements to recreate the heritage site in virtual space combined with ambient sounds. Throughout this paper, we will unpack the elements that contributed to each category in more detail.

This paper demonstrates how intergenerational codesign not only enriches the engagement of older adults with immersive technologies but also crucially informs the development of accessible and intuitive user interfaces in cultural heritage and underwater environments. By integrating extensive user feedback into the design process, we offer novel insights into creating more inclusive and effective immersive experiences.

The remainder of the paper is structured as follows. Section II provides a background and overview of our approach to codesign in the ICONIC project including the exploratory aspect of technology development. Section III describes the methodology and approach in the implementation of the project with a focus on contributors recruitment, partner involvement, and the iterative development process that integrates feedback from the codesign team. Section IV explores the outcome for each package, focusing on the immersive application for each package, and highlighting additional recommendations and findings. Section V unpacks the findings focusing on the HXR and UT rankings based on their respective design priorities. Section VI provides a discussion of common elements in packages and also highlights some of the differences between them. Section VII draws the conclusions of the article and highlights future work.

II. Background

This project approaches codesign from a participatory perspective, accessing the dormant creative potential of intergenerational groups through hands-on creative methods including storytelling and experience design, technology development,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

and interaction design. Here, the project has followed precedent from prior research [9] suggesting ten principles to codesign XR experiences for health interventions in rural communities [10], allowing end users to envision future and speculative scenarios (in this case, shaped by engaging with a local heritage location), the delivery of a research-in-residence approach [11], and the contextualization of XR design within defined societal groups and geographies, among others. These guidelines stem mainly from the GOALD project [12], which focused on physical activity and reminiscence and included a menu of XR experiences for further evaluation and feedback from older adults groups. In addition to those results emerging from the participation of older adults, we pose the question of whether rural communities can benefit from the creative participation of young people in need of high-value digital skills. Furthermore, the literature [13] [14] indicates that youth disenfranchisement results in more critical social outcomes, e.g., vulnerability, mental health and isolation, and feelings of 'nowhere to go'.

Heritage sites and museums play a vital role in the lives of many individuals, particularly in rural communities across South-West England. A 2018 UK Government report examined the influence of historic sites on wellbeing and one of the highlights is the concept of "Heritage as Place" [15]. The report, emphasises the importance of belonging, where the connection to historic locations contributes to reduced social isolation and strengthens feelings of pride, identity, and community. However, these advantages are not experienced equally across all segments of society, with older generation visitors experiencing accessibility and mobility issues. From a large partner network of 36 organisations [16], we partnered with Cotehele - National Trust to create a digital immersive version of the heritage site. Cotehele is a historic estate located in Cornwall, England, and features a medieval manor house, gardens, and a mill. The Great Hall, a significant architectural point of interest of Cotehele, was constructed in the late 15th century and has been preserved in traditional Georgian style for centuries.

With Ocean Conservation Trust (OCT), Plymouth, England, as a partner for the UT work package, the ICONIC project explores how these immersive tools can bridge physical and financial barriers, providing access to the otherwise inaccessible underwater world. Inspired by the beneficial effects of blue spaces on wellbeing [17], our goal is to simulate an underwater experience, enabling people onshore to explore marine environments that are otherwise out of reach. The codesign workshops centered on designing an interaction with the marine environment to evoke a sense of "being underwater" while addressing the practical challenges associated with such a design. The concept of telepresence — originally introduced in human-computer interaction to describe the illusion of being present in a distant location — has evolved beyond its technical roots. Sheridan [18] describes telepresence as the sensation of being "there" at a remote site. Within the scope of the ICONIC project, we have adopted a broader phenomenological view emphasising the sense of presence in underwater spaces that could include methods such as VR and pre-recorded content, enabling engagement without real-time telecommunication [19].

In line with design thinking, we have started the development from the first principles, by understanding what is it that stands out the most about the heritage sites and underwater spaces to our codesign participants and what are the main barriers to higher levels of engagement with these spaces. The goal of the project was to allow contributors to explore different modalities of immersion and interaction and to allow them to define what it means to experience a heritage or underwater spaces in an immersive environment. The technical solutions differ in a way they are delivered and in the resulting level of accessibility, interactivity and immersion for both packages.

Figure 2. During the codesign session, participants prototyped immersive interactions by annotating paper templates of 360° environments, taking snapshots, and instantly exploring their work in Google Cardboard VR. This iterative approach promoted deeper understanding and collaboration.

III. Methods

The ICONIC project recruited twelve Digitally Excluded Older People (DEOP) aged 50+ for the HXR workshop group, and a further twelve DEOP for the UT workshop group. Six Young People (YP) were recruited for the HXR, and an additional six YP were recruited for the UT group. Attendance amongst the younger cohort for both groups was more inconsistent than amongst DEOP contributors. To support asynchronous codesign an additional group of 20 YP were recruited through a higher education partner of the ICONIC project. The project received approval from the Ethics Committee of the University of Plymouth and each contributor received an information sheet and offered the option to withdraw from the study at any time. After each workshop, the contributors were encouraged to raise any issues or provide feedback in person and anonymously through a suggestion box. For analysis, data was anonymised and kept secure on the University of Plymouth machine and OneDrive protected by passwords. Access to the data was limited to the ICONIC team. Codesign

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

participants and partner organisations have given permission for the photos used in this article.

The development process was created to reflect the contributors' involvement at different stages of the codesign process. The five main stages are as follows: Problem Framing, Ideation, Physical prototyping, Digital Design and Testing and Feedback, as described in Figure 1.

Each workshop followed a four-step iterative design approach as described by Macklin and Sharp [20]: Conceptualise, Prototype, Playtest and Evaluation. In the Concept phase - The research team will generate the concepts that it wanted to explore next based on the current development stage of the application and the feedback received. Prototype phase - the concepts get transferred into codesign activities and a technical prototype gets created. The playtest phase - during the workshop the codesign team generates feedback and knowledge through testing the prototypes and executing the designated activities. Evaluation phase - after the workshop the research team evaluates the workshop results, both from activities and the prototype feedback, and generates a new set of concepts to explore for the next workshop. This approach was applied to both the HXR package and the UT package as described by Jones et al. [16]. There were similarities in data collection between the two work packages, as the focus of each was understanding contributors' needs and design priorities for the two technologies. Each workshop featured a variety of activities that were designed to produce written or verbal feedback to support the iterative design of the technologies (Figure 2). Workshop materials and audio recordings were cleaned and transcribed verbatim for thematic analysis.

With Cotehele as a partner for the HXR, contributors chose from a set of six possible indoor locations part of the Cotehele's manor by using 360-degree videos captured at each location. The codesign team selected the Great Hall (Figure 3) due to its impressive size and extensive range of historical artefacts on display, although other locations, such as the kitchen, had great potential in exploring novel and immersive interactions.

The immersive nature of a VR experience is not only limited to the content presented to the user but also how the user interacts with the content. Lack of interaction creates a world where the user feels detached with no real connection to the digital environment, a passive experience where the user is "looking at" something as opposed to an active experience where the user feels "being part" of something [21]. In the HXR, the lack of interaction was clear from the feedback received during the early stages of development with participants exploring 360 videos of various rooms in Cotehele. Statements such as "lack of interaction" or "Frustration over lack of interaction" were provided as feedback. In addition, when participants were asked to complete the sentence "What I wish…" the top priority was "Immersive interaction". The single perspective of a 360-video capture affected some of the participants sense of embodiment, especially when the height of the user didn't match the height of the capture point, we received feedback such as "Feeling of floating" or "Disembodiment". And finally

Figure 3. Image taken inside the Great Hall in Cotehele - National Trust UK.

lack of mobility inside the video was another issue highlighted by the participants through statements such as "Lack of movement" and for the question "What I wish…" the answer was "Movement". These early findings highlighted the importance of interaction, movement and multisensory engagement for the older adults and informed us that the historical content is not enough to deliver an immersive experience.

For the UT, as contributors prioritized local marine environments early on, we focused on prototyping the interaction with the footage from two National Marine Aquarium (NMA) Plymouth tanks dedicated to local fauna and flora: Plymouth Sound and Eddystone reef. The prototype leveraged 360-degree camera footage, Oculus Quest headset [22] and artificial intelligence (AI) for interactive species identification (implemented as an OpenAI API placeholder for the time being). A point-and-click interaction method was selected based on contributors' feedback, with the option to scroll through the menu and the collection of species. The features of the prototype make it suitable for deployment in care homes, schools, and even tourism hubs, offering a scalable model for broad outreach and engagement.

Towards the end of the codesign process for both technologies, participants individually ranked design priorities for the HXR and UT technologies. When undertaking the ranking exercise, participants were asked "What features need to be priotisised for future design of this technology?" Design priorities were each derived from qualitative data collected over the course of codesign workshops. For the HXR technology, the eight design priorities comprised:

- Educational: The VR experience provides the user with detailed information about Cotehele's heritage
- Interactive: Lets users engage with artifacts and historical environments hands-on
- Explorable: Allows users to freely explore Cotehele
- Accessible: Designed for users of all abilities and tech-

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Figure 4. On the left, HXR - final version of the Cotehele's Great Hall aiming to capture the unique "look and feel" from inside the great hall. On the right, UT. An example of interaction with the 360-degree footage of the underwater environment.

nological expertise
- Entertaining: Uses the history of Cotehele to provide a fun experience for users
- Immersive: Provides the user with a sense of being at Cotehele
- Social: Allows multiple users to explore Cotehele in the same virtual space
- Customisable: Provides the user with options to tweak the experience to suit their needs

The eight design priorities for the UT technology comprised:

- Accessible: Designed for users of all abilities and technological expertise
- Educational: The technology provides the user with information about the underwater enviornment
- Entertaining: Uses the underwater space to provide a fun experience for users
- Explorable: Allows users to freely explore an underwater space
- Immersive: Provides the user with a sense of being underwater
- Interactive: The user can engage hands-on with the underwater environment
- Relaxing: Provides a passive, therapeutic experience for the user
- Social: Allows multiple users to explore the underwater space at the same time

To gain a wider understanding of the design priorities for both technologies, participants that were not in the original HXR and UT codesign groups were asked to rank the design priorities. This would have the potential to validate the codesign observations from both HXR and UT workshop groups, should the rankings reflect the observations described previously. In total, the ranking was conducted by $N = 57$ participants (13 from the original HXR codesign group, 44 from additional codesign groups recruited for the ICONIC project) for the HXR technology and $N = 53$ participants (11 from the original UT codesign group, 42 from additional codesign groups recruited for the ICONIC project) for the UT technology.

IV. Codesign Outcomes

Although both packages have an immersive experience as the primary outcome, there are other secondary aspects that emerged as a result of the codesign workshops.

A. Heritage XR

1) Multimodal immersive experience: The main outcome of the HXR package is an immersive VR experience replicating the Cotehele's Great Hall (see the image on the left in Figure 4). The application is a multisensory experience that makes use of visual, audio and haptic feedback. The VR is delivered using the Quest 2 headset developed by Meta [23], which includes two VR controllers. The headset features six degrees of freedom (6DoF) using an inside-out tracking system and is equipped with a set of speakers that allows the delivery of 3-dimensional sound. The tracked controllers support 6DoF and have customisable buttons and haptic feedback capabilities. The virtual space has been created using a combination of local textures, rough measurements, and recreation of the main features of the hall. A set of 4 unique historical artefacts have been scanned using Photogrammetry [24] and due to limited resources and time constraints, the rest of the artefacts in the Great Hall are 3D digital replicas of weapons and items acquired through the Unreal Engine asset store [25].

2) Technical prototype with simplified interactions aimed at older adults: The codesign process revealed the challenges older adults face with various metaphors of interaction, found in most VR applications. Therefore, a set of simplified interactions were created aimed at alleviating some of the issues, such as holding a button pressed for a long time or hard-to-reach buttons. A combination of button mapping and interactive

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

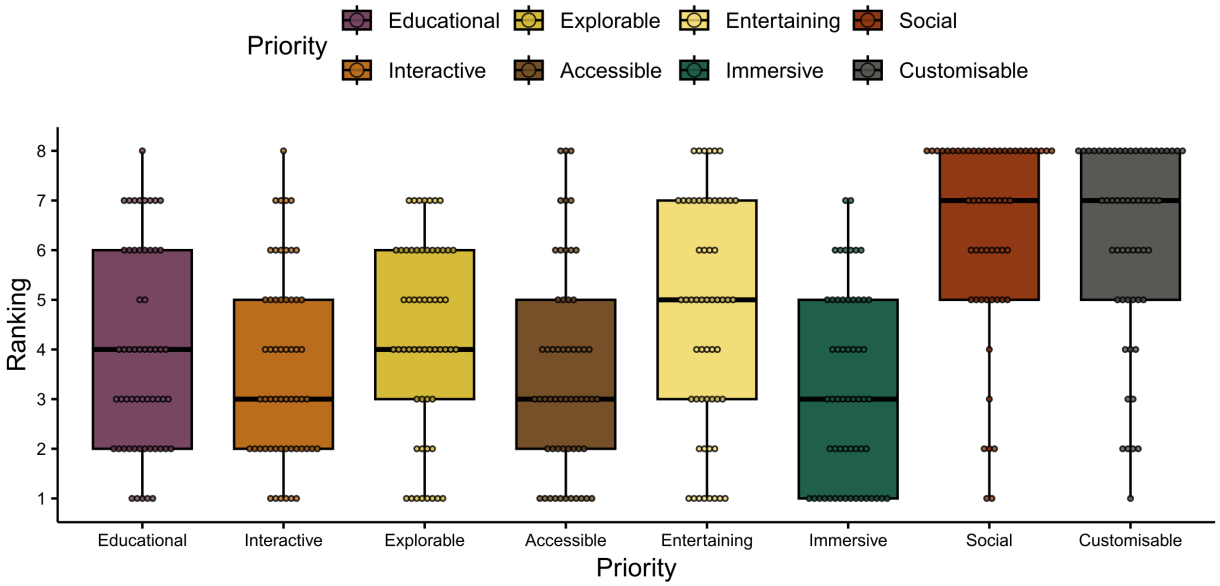


Figure 5. This graph depicts median and interquartile ranges of the HXR rankings for all participants. Each indiivdual point represents one participant's ranking score.

objects was packaged in an example project in Unreal Engine [26]. The interactive elements are modular and flexible, and developers can turn any asset into an interactive object.

3) UX recommendation for the development of VR applications: A set of recommendations for the development of a technical immersive application for older adults through an intergenerational codesign approach. These recommendations are in the process of being published soon in a peer reviewed article.

B. Underwater Telepresence

There are three outcomes from the UT codesign sessions:

1) Immersive Prototype: The main outcome is the immersive VR prototype designed to include most of the features designed by our contributors. There are two modes of interaction: "learning" and "relaxation". Relaxation is designed for users seeking a calming experience; this mode emphasizes the serene beauty of the underwater world with ambient sounds and minimal distractions. The learning (or stimulating) mode enables users to interact and identify marine species within the immersive environment implemented in the prorotype via API calls to a vision enabled LLM system. Features such as "collecting" fish and a virtual agent, designed as a friendly "penguin," engages users with contextual challenges (such as questions about the previously envountered marine species) to encourage users to explore the space more actively. Feedback emphasized the need for realistic, but not necessarily real sound, with ambient underwater sounds enhancing immersion.

2) Design of Alternative Delivery Modes: While there was general agreement about the use of the VR headset, the contributors proposed additional modes of delivery of the experience. The proposed design included a web interface with the ability to interact with 360-degree underwater footage in the same way as before, use of a large interactive screen (that was used to demonstrate the live video in Workshop 2), emphasising the social interaction aspect as an important part of the design. One of the alternative designs to the headset included a portable mini-dome. While the dome enabled shared experiences, it was deemed to be less immersive compared to headsets and limited in scalability due to the infrastructure required. In all design decisions, contributors prioritised accessibility and scalability above most other design properties, thus resulting in the VR headset (Oculus Quest 3) prefered as a final delivery channel.

3) Established Feasibility of Using ROV for Outreach: Remote Operated Vehicles (ROVs) emerged as a promising outreach tool, offering a hands-on, interactive modality to experience underwater environments. We tested the feasibility of the ROV-based telepresence project in an outdoor setting by using BlueROV2, allowing contributors to directly engage in marine exploration. The contributors operated the ROV, navigating an underwater outdoor space in real-time. To improve the comfort of contributors, we have also offered one-to-one ROV teleoperation training sessions in the indoor pool. Throughout the workshop, contributors expressed interest in extending interactivity through robotic arms for activities such as object collection or habitat observation. While not directly related to the immersive prototype, it shows that the ability to control movement in the underwater space can serve as equally engaging immersive modality as high resolution immersive footage. It was decided to continute develop this as an independent thread in the context of Plymouth National Marine Park activities.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

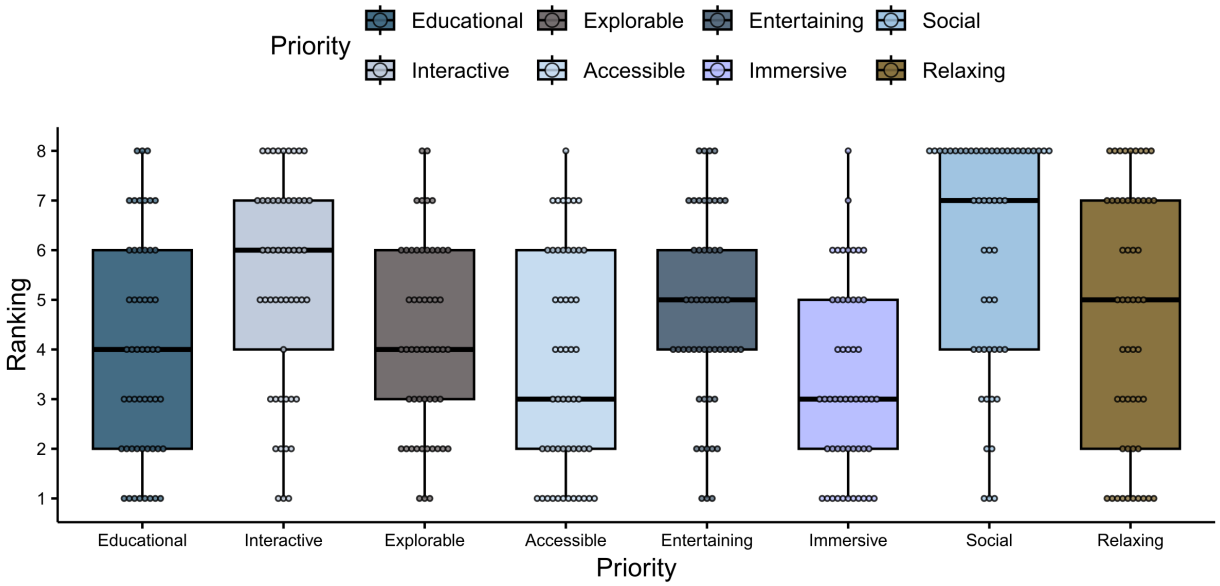


Figure 6. This graph depicts median and interquartile ranges of the UT rankings for all participants. Each indiivdual point represents one participant's ranking score.

C. Analysis

The ranking data for each technology were analysed using R [27] in RStudio [28]. To identify if there were differences in rankings between the main codesign group for each technology and the additional codesign groups that conducted the ranking exercise, Wilcox tests with Bonferroni multiple comparison corrections were conducted between the two groups for each design priority. For both technologies, there were no significant results from the Wilcox tests (all *p*s ¿ .005) so the data from all participants was aggregated into a single dataset. Friedman rank sum tests were conducted on the ranking data to determine if there were differences in how the constructs were ranked. These were followed by Nemenyi post-hoc tests with multiplen comparison corrections to determine if constructs differed in their ranking. Additionally, Kendall's coefficient of concordance W tests were conducted to determine if there was concordance within the rankings across participants.

V. Results

The HXR rankings are visualised in Figure 5. The Friedman rank sum test was significant, $\chi^2(7) = 100.5$, *p* ¡ .001. The Nemenyi post hoc tests indicated that the Social priority was ranked significantly lower than the Educational, *p* ¡ .001, Interactive, *p* ¡ .001, Explorable, *p* ¡ .001, Accessible, *p* ¡ .001, Entertaining, *p* = .004, and Immersive, *p* ¡ .001, design priorities. Similarly, the Customisable design priority was ranked lower than Educational, *p* ¡ .001, Interactive, *p* ¡ .001, Explorable, *p* ¡ .001, Accessible, *p* ¡ .001, Entertaining, *p* = .015, and Immersive, *p* ¡ .001. The only other significant difference was that Immersive was ranked higer than Entertaining, *p* = .023. All other pairwise comparisons were not significant (*p*s ¿ .005)

The Kendall's coefficient of concordance W for the HXR rankings was not significant, $\chi^2(56) = 4.43$, *W* = 0.01, *p* ¿ .005, indicating that participants were not consistently aligned with their rankings.

The UT rankings are visualised in Figure 6. The Friedman rank sum test was significant, $\chi^2(7) = 40.94$, *p* ¡ .001. The Nemenyi post hoc tests indicated that the Social priority was ranked significantly lower than Educational, *p* = .037, Explorable, *p* = .037, Entertaining, *p* ¡ .001, and Immersive, *p* ¡ .001. Additionally, the Immersive priority was ranked significantly higher than Interactive, *p* ¡ .001, and Entertaining, *p* = .037, and the Accessible priority was ranked significantly higher than the Interactive priority, *p* = .006. All other pairwise comparisons were not significant (*p*s ¿ .005)

The Kendall's coefficient of concordance W for the UT rankings was not significant, $\chi^2(52) = 2.3$, *W* = 0.006, *p* ¿ .005, indicating that participants did not all rank the constructs similarly.

VI. Discussion

The generalist approach taken in the development of HXR and UT offered a unique opportunity to explore multiple directions and delivery methods for the experience. Rather than narrowing the scope early, the projects deliberately kept the solution space broad, enabling the team to investigate a variety of technologies and approaches. For HXR, our initial findings identified similar challenges to Wu et al. [29], with older adults experiencing difficulties, such as headset-related neck fatigue and limited field of view leading to extra head movement leading to decreased motor performance. Therefore, the interaction and locomotion were prioritiesed by the codesign team to improve accessibility and direct interaction.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

For the UT the focus was on exploring various methods of allowing contributors to experience telepresence through controlled ROVs and recorded underwater environments in VR. By allowing the contributors to interact with a full spectrum of possible telepresence delivery systems (live stream vs pre-recorded, flat vs 3D, simulated va real, etc.) we were able to identify what aspects of the experience users value most with great level of precision. However, this openness also posed challenges, in particular, highlighting the need for diverse expertise and ample resources to deliver multitude of prototypes with required level of functionality.

Accessibility emerged as a decisive factor in nearly every stage of the project. Whether considering the mobility of devices, their cost, or ease of use, the contributors consistently emphasized the importance of making the technology as inclusive as possible. For instance, while immersive dome projections offered an interesting social dynamic, they were ultimately de-prioritized due to their high infrastructure requirements and limited portability. For both packages, the Quest 2 headset (upgraded to Quest 3) was chosen by the contributors for its affordability, accessibility and mobility. These decisions reinforced the need to prioritize technologies that could reach the largest audience, even if it meant compromising certain experiential features.

Figure 7. We replaced the original elasticated strap for the headset with a more mechanical strap that distributes the weight of the VR headset equally.

A. Exploration through technology interaction

The ability to explore through interaction was one of the key factors that the codesign team highlighted early in the process in both the HXR and UT packages. Workshop 1 was dedicated to problem framing, the contributors explored technologies dedicated to each package in order to identify the key elements of an immersive heritage and underwater telepresence experience.

For HXR, some key elements highlighted by the team were: (i) accessing heritage information through novel and unique ways such as a non-player character (NPC) that provides information about the various historical artefacts; (ii) collaboration between two or more users in exploring the digital space; and (iii) education and knowledge through gamified experiences such as an escape-the-room puzzle.

The exploratory aspect of the codesign process combined with the onboarding sessions and equipment support [30] from the researchers provided contributors with a unique opportunity to explore both the limitations and possibilities inherent in technology, leading to a deeper understanding, reduced cognitive load and reducing their initial reluctance to engage. This was evidenced by the preferences for movement in the VR environment. Initially, teleportation as the initial locomotion metaphor, proved challenging for some of the contributors. A combination of controls and the ability to aim towards a landing spot made some contributors uncomfortable. A more simple locomotion, called ”grab-and-drag” was initially preferred, but as the contributors became more experienced with the VR technology, they started to revert to the teleportation metaphor. The contributors increase in confidence in using the VR assembly shifted the focus towards the creative aspects of engaging with the content, especially in the workshops that took place at a later stage in the process. This is consistent with the findings of Zhang [7], who argue that during technology development, the involvement of older users is crucial, especially if the final goal is the adoption of technology [31].

For UT, one of the most striking findings was the evolution of the preferences of the contributors as they interacted with the technology. For example, while live streaming was initially considered a priority, this changed once contributors received live-like footage (footage recorded earlier). The limitations in video quality led to a diminished preference for live streaming in favour of pre-recorded footage. Similarly, in the first workshop, real underwater sounds were thought to be crucial for immersion, but after interacting with the videos with real underwater sounds recorded by hydrophones, participants found that they did not match their expectation of a relaxing ambience. Hence, we have used an ambient underwater sound in the consequent prototypes.

Another example of this evolution occurred during the ROV trials. Initially, contributors viewed interaction with the underwater environment strictly negatively, as disruptive to the environment and in conflict with the environmental preservation motivations of the group. However, ROV’s teleoperation capabilities introduced a new dimension of engagement, and contributors became enthusiastic about more direct interactivity. The session provided suggestions for features such as a robotic arm to “collect samples” or clean up the ocean floor from debri.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

B. Accessibility, usability and comfort

A variety of immersive experiences have different levels of immersion based on the technology used to deliver them and the implementation of the experience. Immersion is an objective factor in a system that mirrors the extent of technology can support natural sensorimotor alternatives to perception [32]. The contributing factors are typically related to the real world, mainly the hardware specifications or the design of the system, such as resolution, panoramic view, audio input and generally the number of outside physical realities that are blocked by the system [33]. The importance of this was clear from the beginning, in both packages, with members of the codesign team expressing discomfort with the headset. This is all related to the weight distribution of the headset [34] using the original strap, as it adds pressure on the forehead of the user and strain on the neck. This issue has been solved by purchasing a custom head strap that allows weight distribution, transferring the pressure from the forehead to the rest of the strap (Figure 7). Although the new strap marginally increased the overall weight of the system, the contributors reported increase comfort in wearing the VR headset, which in turn, improved the quality of the experience. The new head strap also reduced the amount of light reaching the user's eyes through an improved light blocker that sat closer to the face. Participant comments, such as appreciating the "adjustable headset," and "The headset was a much better fit than previous ones and made it a more comfortable experience" clearly evidenced the positive impact of these ergonomic improvements.

Figure 8. The controller has a strap attached to it that allows the user to open their hand and the controller will stay attached, minimising the amount of strain on the hand.

The use of the VR controller also presented a challenge early in the codesign process. Although ergonomic in its design, the needs and requirements for older adults present different challenges. For starters, holding the controller in hand for a long time presents a challenge, considering the participants might suffer from frailty or from a limited range of motion in the hand. The solution was to add a strap to the controllers that allowed participants to slide the controller on their hand, and the controller was attached to the hand without having to actually close their hands. This greatly reduced the strain on the hand and one participant actually mentioned that it is "easy to control" (Figure 8).

In addition to the ergonomic aspect of the controller, participants complained about the controllers, "difficult to get used to" or "Not being able to use the controls with confidence" or even "The controls - they were not intuitive". The complexity of using the controller came from various buttons and joysticks that had different actions attached to them such as teleporting, moving, grabbing objects, opening various options and also from the locations of some of the buttons that were difficult to reach by some of the users. In response to that we simplified the controls to use only one button, the trigger button. The button was easy to reach, and we attached additional behaviours to the button depending on the context. All the actions were toggable as participants were able to grab objects close or from a distance by using just one press of the button. On the second press the button disable the grab action and the object fell to the ground. That means the participants didn't have to keep the button pressed in order to hold an object in their hands.

C. Immersive experience for older adults

Focusing on the digital content is paramount to immerse the player in the digital environment and gives the user a sense of presence (SoP), the sensation of leaving their current location, and they transport to a virtual environment where they act as if they are physically there, perceiving virtual objects and individuals as real [33] [35] [36]. For the HXR, in our approach to increase the SoP for our contributors, we introduce multisensory inputs using visual, audio and touch. Many of the historical artefacts are interactable, with the user having the ability to grab them from proximity or from a distance (Figure 9).

In the early stages of the codesign process, after the transition from a 360-video to an immersive interactable version of Cotehele participants expressed their enthusiasm and appreciation for interaction such as "picking things up" and "actually handle things". These early interactions used objects of various shapes such as cubes and a gun that shoots small balls. The goal was to explore the grabbing mechanic involving the hand/controller in VR and the object and also to explore how the objects behave in the environment. The participants used a set of controllers that were replicated in VR to align their real-world sensory inputs to the visuals in the VR environment. The visual representation of the grab mechanic was done through replacing the controller with the

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

grabbed object with positive feedback from the participants "I loved picking up the gun and firing it".

We also explored the realistic behaviour of simple objects such as cubes in order to increase the participant's sense of presence in the virtual environment. We introduced physics-based cubes that mimics the real behaviour of a cube in the real world. Participants had the ability to grab, move, push or throw these cubes with positive comments such as "the cubes are very realistic and fall/move as they should" further supporting the objective factor of Immersion. Of course, in the later codesign stages we introduce actual interactive historic artefacts and elements that contribute to the narrative of the experience but these early stages prototypes played a crucial role in understanding the codesigners preferences and needs.

Figure 9. The interaction works for nearby objects or for objects in the distance. The feedback is in the form of visual highlight, haptic feedback and audio for when the object lands in the user's hand.

Audio textures are used for simple interactions or for impacts between swords and other objects in the environment in order to give user situational awareness [37]. These interactions are accompanied by haptic feedback in the form of small vibrations with different amplitudes and intensities in order to trigger tasks [38] and enhance the level of immersion for the user. Multisensory interaction was one of the codesign group's priorities with an initial discussion about implementing hand tracking in order to simplify the interaction metaphor versus keeping the controllers with a simplified version of the button mapping. The contributors opted for continued use of the controllers as they did not want to lose the haptic feedback.

For the UT, immersion was embodied through a bimodal distribution. Many contributors expressed interest in highly stimulating and relaxing experiences, depending on the context. For example, some valued the calmness and meditative quality of simply observing marine life, while others were drawn to active gamified elements that encouraged exploration and learning. This dual demand for contrasting modes presented a design challenge, but also highlighted the potential versatility of the system by catering to diverse user needs. We have responded by delvering two modes of interaction, and found that both were equally used by the participants. Some participants found it is be uncomfortable to be "inside" of the tank in the VR space, without any visible support underneath them, making them loose balance. It suggests that building some aditional virtual structure around the user could be beneficial to improve this experience.

The codesign exploratory approach of the ICONIC project was aimed at creating a technological base for each of the four technologies, with the intention of one or more social enterprises to take over the development and turn each prototype into a product developed by the local community for the local community. This approach meant that we did not run any simulator sickness tests [reference] with our codesign groups, although we encourage them at each session to report any symptoms.

D. Future design priorities

To identify the future design priorities for the two technologies participants from our codesign workshops and from separate evaluation workshops ranked eight design priorities derived qualitatively from codesign data. Between the HXR and UT workshop data, the design priorities were identical, with the only difference being a 'Customisable' design priority for the HXR experience and a 'Relaxing' design priority for the UT technology.

Analyses of the HXR ranking data indicated that participants did not demonstrate concordance in their overall ranking patterns. There were, however, significant differences between the ranking of specific items. Participants significantly prioritised immersion, interaction, and accessibility to a greater extent than social or customisable elements. These reflect observations from the HXR codesign workshops as participants sought to maximise the immersive nature of this technology. The emphasis on accessibility also reflects difficulties some older participants experienced with Quest 3 controllers, highlighting the importance of inclusive design in VR systems and software.

Similarly to the ranking data for the HXR technology, there was no concordance amongst participant rankings for the UT technology, though, again, there were significant differences between specific priorities. Participants ranked the social design priority lower than the educational, explorable, entertaining, and immersive design priorities. As with the HXR codesign process, these data reflect observations from the UT codesign workshops.

While rankings across the two technologies focused on different sets of priorities, both consistently deprioritized social features relative to technology-specific core functionalities, i.e., Immersion, Accessibility, and Interactivity for the HXR technology, and Education, Exploration, and Entertainment for

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the UT technology. This deprioritisation of social elements of the technology may reflect the stage of the codesign process, as the features that were prioritised reflect the further developmeent of core functionality that is integral to the identity of the two technologies. When the core functionality is well-established, it may be that participants then begin to prioritise elements that were not a clear design priority at this stage. Whilst these data do not demonstrate a concordance for design priorities across the two technologies, it does highlight priority areas of development for the two technologies that should be prioritised, and others that should be deprioritised.

VII. Conclusion and Future Work

In conclusion, the comparative analysis of design priorities across the HXR and UT technologies highlights both shared and technology-specific user expectations that can inform future development of immersive heritage and underwater telepresence experiences. Despite limited concordance in participant rankings, consistent patterns emerged that underscore the centrality of immersion, accessibility, and interactivity as critical design priorities for HXR, and education, exploration, and entertainment for UT. The consistent de-prioritisation of social features across both technologies may reflect the developmental stage of the prototypes, where participants' focus remained on refining core functional elements essential to meaningful engagement before contemplating social engagement.

These findings reinforce the importance of iterative, user-centred design processes in the evolution of immersive technologies for older adults. Through an iterative development process and a series of intergenerational codesign workshops, the research demonstrates that immersive XR technologies hold considerable potential for reconstructing cultural heritage sites and underwater environments in ways that are both engaging and meaningful for older users. Nevertheless, the effective adoption of these technologies remains dependent on addressing key design challenges, including the implementation of intuitive locomotion and interaction mechanisms, the simplification of control schemes, and the incorporation of ergonomic solutions to enhance comfort and mitigate physical strain. Notably, the introduction of simplified locomotion and control systems instilled greater user confidence in operating the VR headset, enabling participants to shift their focus from issues of usability and hardware management toward a more creative and exploratory engagement with heritage artefacts.

Moreover, factors such as accessibility, affordability, and hardware usability emerged as critical priorities for older participants, underscoring the necessity of minimizing barriers to engagement. To address affordability constraints, the codesign team selected a cost-effective headset, though its limited technical capabilities posed challenges in achieving high-fidelity visual representation. In response to usability and accessibility concerns and as a result of collaborative design interventions, simple non-expensive solutions were introduced to enhance comfort and ease of use—these included the integration of adjustable straps to secure the controllers to the users' hands and a redesigned head strap that more evenly distributed the device's weight. Furthermore, a context-sensitive single-button interface was developed to adapt dynamically to user actions within the virtual environment, thereby simplifying interaction and reducing cognitive load. Collectively, the results provide actionable insights for guiding the next phase of design and evaluation, advancing the creation of inclusive, engaging, and contextually resonant XR experiences for diverse audiences.

The value of codesign became particularly evident in the interactions contributors had with multiple version of the technology prototypes. Participants initially articulated preferences and expectations based on their imagined interactions with the technology. Once these ideas were translated into prototypes and participants could directly interact with them, their understanding frequently evolved. Experiencing the designs in a tangible form allowed participants to reassess their assumptions, refine their preferences, and, in some cases, develop entirely new ideas. This iterative process demonstrated that meaningful codesign extends beyond conversation. The designs emerge through engagement, reflection, and learning that occur when participants interact with evolving design artifacts. Such embodied participation grounds the technology's development in lived experience.

Ongoing evaluation sessions with intergenerational user groups and industry partners are currently underway to validate and extend the findings of this study. The insights emerging from these sessions will inform the next phase of development, with a particular focus on designing and testing a simplified controller that enhances usability, supports personalisation, and reduces cognitive load for older adults with limited mobility. This next stage will further advance the project's commitment to inclusive design, aiming to translate user-centred insights into tangible technological solutions that foster broader accessibility and engagement across generations.

Acknowledgment

This paper is presented on behalf of the ICONIC project that includes Katharine Willis, Daniel Maudlin, Chunxu Li, Sheena Asthana, Kerry Howell, Shangming Zhou, Emmanuel Ifeachor, and Hannah Bradwell as co-applicants and advisors and Lauren Tenn (Media and Administration Officer). Intergenerational co-creation of novel technologies to reconnect digitally excluded people with community & cultural landscapes in coastal economies (ICONIC) is funded by UK Research and Innovation Engineering and Physical Sciences Research Council Grant Ref: EP/W024357/1.

We thank Ocean Conservation Trust, Cotehele National Trust, Marine Station staff, Oliver Smith and Benjamin Buckfield-Richards for their contribution to this project.

References

[1] M. Varga et al., 'Intergenerational codesign of immersive technology for a heritage site and underwater VR experience,' in The Second International Conference on Artificial Intelligence and Immersive Virtual Reality (AIVR 2025). IARIA, 2025.

[2] 'AIVR 2025,' The Second International Conference on Artificial Intelligence and Immersive Virtual Reality, Accessed: Nov. 25, 2025. [Online]. Available at: https://www.iaria.org/conferences2025/AIVR25.html.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[3] M. N. Varga, R. Merrison-Hort, P. Watson, R. Borisyuk, and D. Livingstone, 'Tadpole VR: Virtual reality visualization of a simulated tadpole spinal cord,' Virtual Reality, vol. 25, no. 1, pp. 1–17, Mar. 2021, ISSN: 1434-9957. DOI: 10.1007/s10055-020-00431-z.

[4] M. N. Varga, S. E. Gaudl, and D. J. Walker, 'Many-objective population visualisation with geons,' Proceedings of the Genetic and Evolutionary Computation Conference Companion, ser. GECCO '21, Lille, France: Association for Computing Machinery, 2021, pp. 1961–1969, ISBN: 9781450383516. DOI: 10.1145/3449726.3463162.

[5] M. N. Varga and D. Bazazian, 'Interactive Heritage Site Mobile Application on Artworks,' in Advances in Representation: New AI and XR-Driven Transdisciplinarity, A. Giordano, M. Russo, and R. Spallone, Eds., Cham: Springer Nature Switzerland, 2024, pp. 125–139, ISBN: 978-3-031-62963-1. DOI: 10.1007/978-3-031-62963-1_8.

[6] E. Ch'ng, Y. Li, S. Cai, and F. T. Leow, 'The Effects of VR Environments on the Acceptance, Experience, and Expectations of Cultural Heritage Learning,' J. Comput. Cult. Herit., vol. 13, no. 1, pp. 1–21, Feb. 2020, ISSN: 1556-4673. DOI: 10.1145/3352933.

[7] M. Zhang, 'Older people's attitudes towards emerging technologies: A systematic literature review,' Public Understanding of Science, vol. 32, no. 8, pp. 948–968, Nov. 2023, ISSN: 0963-6625. DOI: 10.1177/09636625231171677.

[8] K. Ijaz, T. T. M. Tran, A. B. Kocaballi, R. A. Calvo, S. Berkovsky, and N. Ahmadpour, 'Design Considerations for Immersive Virtual Reality Applications for Older Adults: A Scoping Review', Multimodal Technologies and Interaction, vol. 6, no. 7, Art. no. 7, July 2022, doi: 10.3390/mti6070060.

[9] A. V. Reyes, M. N. Varga, H. Bradwell, and R. Baxter, 'Unlocking Social Innovation in XR for Healthcare in Coastal Communities,' IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Mar. 2023, pp. 128–131. DOI: 10.1109/VRW58643.2023.00032.

[10] A. Veliz Reyes, M. N. Varga, H. Bradwell, R. Baxter, R. B. Jones, and D. Maudlin, 'Codesign principles for the effective development of digital heritage extended reality systems for health interventions in rural communities', CoDesign, vol. 20, no. 2, pp. 295–312, Apr. 2024, doi: 10.1080/15710882.2024.2349588.

[11] F. Gradinger, J. Elston, S. Asthana, S. Martin, and R. Byng, 'Reflections on the Researcher-in-Residence model coproducing knowledge for action in an Integrated Care Organisation: A mixed methods case study using an impact survey and field notes,' Evidence & Policy, vol. 15, no. 2, pp. 197–215, May 2019. DOI: 10.1332/174426419X15538508969850.

[12] A. C. Whittaker et al., 'Intergenerational co-production to increase physical activity and social connectivity for health in the GOALD project (Generating Older Active Lives Digitally)', Eur J Public Health, vol. 33, no. Supplement_1, p. ckad133.097, Sept. 2023, doi: 10.1093/eurpub/ckad133.097.

[13] L. Wareing, N. Dunn, and P. A. Rodgers, 'Levelling up north of england for rural youth: Co-design possibilities and challenges,' Blucher Design Proceedings, vol. 9, no. 5, pp. 506–518, 2021, ISSN: 2318-6968. DOI: http://dx.doi.org/10.5151/ead2021-157.

[14] N. Y. Agency, 'Overlooked: Young people and rural youth services,' National Youth Agency, Tech. Rep., 2021, pp. 1–20.

[15] G. O. for Science, 'Mental capital and wellbeing: Making the most of ourselves in the 21st century,' UK Government Office for Science, Oct. 22, 2018.

[16] R. Jones et al., 'Addressing Digital Exclusion via the Inter-Generational Codesign of Extended Reality, Underwater Telepresence, Social Games, and Voice AI Technologies,' International Journal On Advances in Intelligent Systems, vol. 17, no. 3&4, Nov. 2024, ISSN: 1942-2679.

[17] J. Grellier et al., 'BlueHealth: A study programme protocol for mapping and quantifying the potential benefits to public health and well-being from Europe's blue spaces,' BMJ open, vol. 7, no. 6, e016188, Jun. 2017, ISSN: 2044-6055. DOI: 10.1136/bmjopen-2017-016188.

[18] T. B. Sheridan, 'Musings on Telepresence and Virtual Presence,' Presence: Teleoperators and Virtual Environments, vol. 1, no. 1, pp. 120–126, Feb. 1992. DOI: 10 . 1162 / pres.1992.1.1.120.

[19] O. Hagen, A. Aly, R. Jones, M. Varga, and D. Bazazian, 'Beyond the surface: A scoping review of vision-based underwater experience technologies and user studies,' Intelligent Marine Technology and Systems, vol. 2, no. 1, p. 19, Jun. 2024, ISSN: 2948-1953. DOI: 10.1007/s44295-024-00034-3.

[20] C. Macklin and J. Sharp, 'Games, design and play : a detailed approach to iterative game design,' eng, First edition. Boston, MA ; Addison-Wesley, 2016, ISBN: 0134392078.

[21] C. Ferguson, E. L. van den Broek, and H. van Oostendorp, 'On the role of interaction mode and story structure in virtual reality serious games,' Computers Education, vol. 143, p. 103 671, 2020, ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2019.103671.

[22] 'Meta Quest 3,' Quest 3, Accessed: Nov. 23, 2025. [Online]. Available: https://www.meta.com/gb/quest/quest-3/.

[23] 'Meta,' Meta, Accessed: Mar. 1, 2025. [Online]. Available: https://www.meta.com/.

[24] F. Remondino, 'Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning,' Remote Sensing, vol. 3, no. 6, pp. 1104–1138, Jun. 2011, ISSN: 2072-4292. DOI: 10.3390/rs3061104.

[25] 'Fab,' Fab.com, 2025, Accessed: Mar. 1, 2025. [Online]. Available: https://www.fab.com/.

[26] 'Unreal Engine,' Unreal Engine, 2025, Accessed: Mar. 1, 2025. [Online]. Available: https://www.unrealengine.com/en-US/home.

[27] R Core Team, 'R: A language and environment for statistical computing,' R Foundation for Statistical Computing, Vienna, Austria, 2025.

[28] Posit team, 'Rstudio: Integrated development environment for r,' Posit Software, PBC, Boston, MA, 2025.

[29] Z. Wu, D. Wang, S. Zhang, Y. Huang, Z. Wang, and M. Fan, 'Toward Making Virtual Reality (VR) More Inclusive for Older Adults: Investigating Aging Effect on Target Selection and Manipulation Tasks in VR', in Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, in CHI '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–17. doi: 10.1145/3613904.3642558.

[30] S. Baker et al., 'Interrogating Social Virtual Reality as a Communication Medium for Older Adults,' Proc. ACM Hum.Comput. Interact., vol. 3, 149:1–149:24, CSCW Nov. 7, 2019. DOI: 10.1145/3359251.

[31] E. Vaportzis, M. Giatsi Clausen, and A. J. Gow, 'Older Adults Perceptions of Technology and Barriers to Interacting with Tablet Computers: A Focus Group Study,' Frontiers in Psychology, vol. 8, p. 1687, Oct. 2017, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01687.

[32] M. Slater, 'Immersion and the illusion of presence in virtual reality,' British Journal of Psychology, vol. 109, no. 3, pp. 431–433, Aug. 2018, ISSN: 0007-1269, 2044-8295. DOI: 10.1111/bjop.12305.

[33] J.-C. Servotte et al., 'Virtual Reality Experience: Immersion, Sense of Presence, and Cybersickness,' Clinical Simulation in Nursing, vol. 38, pp. 35–43, Jan. 2020, ISSN: 1876-1399. DOI: 10.1016/j.ecns.2019.09.006.

[34] J. R. Bruun-Pedersen, S. Serafin, and L. B. Kofoed, 'Going Outside While Staying Inside — Exercise Motivation with Immersive vs. Non–immersive Recreational Virtual Environment Augmentation for Older Adult Nursing Home Residents,' in 2016 IEEE International Conference on Healthcare Informatics (ICHI), Oct. 2016, pp. 216–226. DOI: 10.1109/ICHI.2016.31.

[35] M. Slater, 'Measuring Presence: A Response to the Witmer and Singer Presence Questionnaire,' Presence, vol. 8, no. 5, pp. 560–565, Oct. 1999, ISSN: 1054-7460. DOI: 10 . 1162 /105474699566477.

[36] M. Slater and M. V. Sanchez-Vives, 'Enhancing Our Lives with Immersive Virtual Reality,' Frontiers in Robotics and AI, vol. 3, Dec. 2016, ISSN: 2296-9144. DOI: 10.3389/frobt.2016.00074.

[37] Y.-C. Yang, 'Role-play in Virtual Reality Game for the Senior,' Proceedings of the 2019 7th International Conference on Information and Education Technology, ser. ICIET 2019, New York, NY, USA: Association for Computing Machinery, Mar. 29, 2019, pp. 31–35, ISBN: 978-1-4503-6639-7. DOI: 10.1145/3323771.3323812.

[38] É. Ouellet, B. Boller, N. Corriveau-Lecavalier, S. Cloutier, and S. Belleville, 'The Virtual Shop: A new immersive virtual reality environment and scenario for the assessment of everyday memory,' Journal of Neuroscience Methods, vol. 303, pp. 126–135, Jun. 1, 2018, ISSN: 1872-678X. DOI: 10.1016/j.jneumeth.2018.03.010 pmid: 29581009.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Direct Adaptive Input Matrix Estimation Approaches for Linear Time-Invariant Dynamic Systems with Known Inputs

Kevin Fuentes, Mark Balas, and James Hubbard
Department of Mechanical Engineering
Texas A&M University
College Station, USA
e-mail: {kfuentes | mbalas | jhubbard}@tamu.edu

***Abstract*—Equations Of Motion (EOM) can effectively describe physical dynamics within a prescribed set of assumptions and constraints. However, in many physical dynamic systems, performance can alternate and deteriorate with age, use, or alterations in operational points or environments. The culmination of these factors will be denoted as a health change later in the text. These physical changes can be characterized as alterations in the constitutive constants, which are denoted as mass, stiffness, and damping for a mechanical system, and internal interactions. If these health changes are not taken into account in the modeled EOM, discrepancies may emerge between the physical and model responses. The paper investigates two scenarios: 1) both the true plant and the input matrix experience a form of health change, and 2) only the input matrix experiences a form of health change. The control schemes depend on knowing the true system's input and output states. For case 1: Lyapunov stability proof guarantees internal and external state error convergence to zero asymptotically if the true system experiences health changes within the assumptions and constraints of the proposed control scheme. For case 2: the internal and external state errors converge asymptotically to a neighborhood of zero rather than to zero itself. This behavior results from the presence of a feedback filter in the adaptive input matrix estimation law, which restricts full convergence but enables faster state-error convergence compared to the case without a feedback filter.**

Keywords-Adaptive; Control; Estimation; Plant; Input Matrix.

I. Introduction

This article extends our original *Adaptive 2025* conference paper by considering feedback filters on the adaptive laws: adaptive state and adaptive input matrix estimation laws [1]. In earlier work, a feedback filter was successfully applied only to the adaptive state estimator law [2]. Initial attempts to apply feedback filters to both adaptive laws simultaneously, or to apply a feedback filter to one while omitting it from the other, did not yield promising simulation results. This is likely due to the imbalance in managing plant and input matrix uncertainty, while the feedback filter enables accelerated convergence relative to the non-feedback filtered case. Given the lack of favorable simulation outcomes, the final implementation applied the feedback filter solely to the adaptive input matrix estimator, while the adaptive state-estimation law was omitted. This article presents two proofs: 1) adaptive state and adaptive input matrix estimation, and 2) adaptive input matrix estimation using a feedback filter.

Before discussing the adaptive laws, it is import to understand their need. Equations Of Motion (EOM) can describe the dynamics of the true physical system within a set of assumptions and constraints. The EOM cannot guarantee anything beyond the prescribed conditions. With age, usage, changes in operational points, or environmental elements, the true physical system can undergo degradation. Failing to account for these degradations or health changes resulting from changes in internal interactions or constitutive constants—such as mass, stiffness, and damping in mechanical systems—can lead to an inaccurate depiction of the true dynamics. Equally important, but often overlooked is the potential decline of the system's actuator, which influences how inputs interact with the physical system. A substantial portion of control problems involves the regulation of output errors concerning a given input. Ignoring the health status changes in system dynamics or actuation can result in catastrophic failure if synthesized inputs does not adequately address these changes.

For traditional Luenberger or Kalman-like estimators to be practical, there has to be minimal uncertainty about the system [3], [4]. Unlike Luenberger Estimators, Kalman-like filters are renowned for their ability to account for noise and stochastic variations resulting from sensor or process disturbances, under the assumption that the noise follows a Gaussian distribution centered around zero. However, neither type of estimator is capable of accommodating changes in the health status of the system dynamics or the input matrix.

The sensitivity of Luenberger and Kalman-like estimators to minimal uncertainty regarding system dynamics motivates the development of robustness techniques to address model uncertainty [5], [6]. The control technique presented in this text can manage both plant and input matrix uncertainties. More importantly, it can also accommodate significant changes in system health, as defined in the derivation. This work builds upon our earlier findings, which indicated that only the true-physical plant experiences a health status change, causing changes in dynamics and constitutive constants [7]. In 2022, *Griffith* developed a closed-loop approach for input matrix estimation [8]. This paper explores the scenario in which the plant and the input matrix experience a change in health and is an extension of [1].

The implemented control architecture was designed for a general system and can be applied to any system that meets the assumptions and constraints outlined in the proof. The proof relies on two primary Lyapunov system stability criteria: Strict Positive Real (SPR) and Almost Strictly Dissipative (ASD). For a more formal definition and detailed explanation of SPR

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

and ASD in the context of stability, please refer to [7], [9]. Moreover, since none of the estimated states are fed back to the true system, the estimator can operate without risking harm to the true system. Additionally, the proposed control scheme can be utilized offline and online. However, there can exist practical and numerical limits.

Following the introduction, this paper is divided into additional sections: III. Main Result - Theorem 1, IV. Main Result - Theorem 2, V. Illustrative Example - Theorem 1, VI. Illustrative Example - Theorem 2, and VII. Discussion. The beginning of Section III offers a summary of the derivation process for adaptive state and adaptive input matrix estimation, presenting one of the paper's theorem and control diagram. Sub-sections III-A and III-B provide the assumptions and constraints for both the true and reference systems while laying the foundation for updating the reference model. Sub-Section III-C defines the error states and their dynamics. In the error dynamics, residual terms exist; therefore, error states cannot be guaranteed to converge to zero. To address this issue, the error dynamic states are treated as energy-like terms. Then, an energy-like balance is constructed to remove residual terms, guaranteeing the error state to converge to zero globally as time approaches infinity. This process is detailed in Section III-D and Section III-E. A similar procedure is presented in Section IV for Theorem 2, adaptive input matrix estimation using a feedback filter. Following the derivations, Section V and Section VI present two Illustrative Examples detailing the implementation of the derived control schemes. These examples detail a scenario where the error state takes a relatively long and short time to converge. In particular, the interaction tuning terms $\{\gamma_u, \gamma_u\}$ were left unadjusted. These terms can affect the time at which the error state converges. Finally, the paper ends with two sections: VII. Discussion and VIII. Conclusion.

II. Nomenclature

A	=	True Plant
A_m	=	Model Plant
ASD	=	Almost Strictly Dissipative
B	=	Input Matrix
B_m	=	Input Matrix Model
$\in$	=	Belongs
C	=	Output Matrix
$(\cdot)^\dagger$	=	Conjugate Transpose
e_x	=	Internal State Error
$\hat{e}_y$	=	External State Error
$\hat{}$	=	Estimate
$\forall$	=	For All
L_*	=	Fixed Correction Matrix
γ	=	Interaction Tuning Term
ΔL	=	Variance Matrix
PR	=	Positive Real
SPR	=	Strictly Positive Real
σ	=	Set of Eigenvalues
$\ni$	=	Such that
Re	=	Real
$\exists$	=	There Exists
u	=	Input
x	=	Internal State
y	=	External (Output) State

III. Main Result - Theorem 1

Pertaining to the work being presented, the derived theorem and control laws, shown in Theorem 1 and Fig. 1, are catered to minimizing the internal state error (e_x) to zero between the true-physical system and reference model. This is achieved by accounting for discrepancies in the model plant (A_m) and input matrix (B_m), given a known input (u), output matrix (C), and external state (y). Uncertainty or variability in the model plant and input matrix means the convergence of the internal state error to zero cannot be guaranteed. As detailed in the derivation, to mitigate any variability, the error system is treated as an energy-like term. The aim is to dissipate all the energy of the error system, thereby ensuring the internal state error converges to zero as time approaches infinity, $e_x \xrightarrow[t\to\infty]{} 0$. To ensure error energy-like dissipation, the energy-like time rate of change for the error system is determined. Subsequently, residual energy-like time rate of change terms from any uncertainty are identified and countered. The remaining energy-like time rate of change term and the use of stability lemma, Barbalat-Lyapunov Lemma, ensures $e_x \xrightarrow[t\to\infty]{} 0$ asymptotically.

Theorem 1: Output Feedback on Reference Model for Adaptive Input Matrix, Plant, and State Estimation.
Consider the following state error system:

$$\begin{cases} \dot{e}_x = (A_m - KC)e_x + B_m\left(\Delta L_1 u + \Delta L_2 y\right) \\ \hat{e}_y = Ce_x = C(\hat{x} - x) \\ L_1 = \Delta L_1 + L_{1*} \\ L_2 = \Delta L_2 + L_{2*} \\ \dot{L}_1 = \Delta\dot{L}_1 = -\hat{e}_y u^\dagger \gamma_u \\ \dot{L}_2 = \Delta\dot{L}_2 = -\hat{e}_y y^\dagger \gamma_y \end{cases}, \quad (1)$$

where e_x is the estimated internal state error, $\hat{e}_y$ is the external estimated state error, $\{L_{1*}, L_{2*}\}$ are fixed-correction matrices, $\{\Delta L_1, \Delta L_2\}$ are the variability-uncertainty terms, K is a fixed gain, and $\{\gamma_u, \gamma_y\} > 0$ are the interaction tuning terms. Given:

1) The triples of (A, B, C) and (A_m, B_m, C) are ASD and SPR respectively.
2) A model plant (A_m) must exist.
3) A model input matrix (B_m) must exist.
4) Output matrix (C) is known.
5) Allow $B \in \text{Sp}\{B_m L_{1*}\} \ni B \equiv B_m L_{1*}$.
6) Allow $A \in \text{Sp}\{A_m, B_m L_{2*} C\} \ni A = A_m + B_m L_{2*} C$.
7) The set of eigenvalues (σ) of the true and reference plant are stable (i.e. $\text{Re}(\sigma(A)) < 0$ & $\text{Re}(\sigma(A_m)) < 0$).

If conditions are met, then $\{e_x, \hat{e}_y\} \xrightarrow[t\to\infty]{} 0$ asymptotically. $\{\Delta L_1, \Delta L_2\}$ are guaranteed to be bounded; however, no guarantee of $\{\Delta L_1, \Delta L_2\} \xrightarrow[t\to\infty]{} 0$. If $\{\Delta L_1, \Delta L_2\} \xrightarrow[t\to\infty]{} 0$, then

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

the dynamics of the true system or some energy equivalence have been numerically captured.

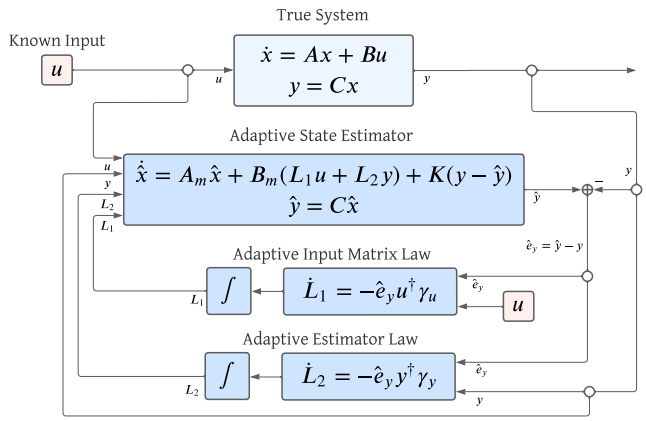


Figure 1. Control diagram for adaptive plant, input matrix, and state estimation given a known input (u), output matrix (C), and external state (y).

A. Defining True System Dynamics

Assume the dynamics of the true-physical system is linear time-invariant and therefore can be expressed in state-space form such that:

$$\text{True System} \begin{cases} \dot{x} = Ax + Bu \\ y = Cx. \end{cases} \tag{2}$$

Both the true system's plant (A), assumed to be stable (i.e. $\text{Re}(\sigma\{A\}) < 0$), and the input matrix (B) experience a health change caused by age or use, altering the constitutive constants and system dynamics. Output matrix (C) and external (output) state (y) are known. The input (u) can be any bounded-continuous waveform the user provides, possibly a known disturbance.

B. Overview of Updating the Reference Model

Subsequent sections will derive a control scheme and laws to minimize the error between the true and reference systems, Eq. (2) and Eq. (3), respectively. Note that both true and model systems match in dimension size, but can differ in constitutive constant values and internal interactions. Allow the reference model to have the following state space characteristics:

$$\text{Reference Model} \begin{cases} \dot{x}_m = A_m x_m + B_m u \\ y_m = Cx_m. \end{cases} \tag{3}$$

To update the input matrix model (B_m), assume that B_m can be corrected via a input matrix fixed correction term (L_{1*}) such that:

$$B \equiv B_m L_{1*}. \tag{4}$$

The true plant is assumed to be decomposed into two components: an initial plant model (A_m) and plant matrix correction term $(B_m L_{2*} C)$ such that:

$$A \equiv A_m + B_m L_{2*} C. \tag{5}$$

Both Eq. (4) and Eq. (5) assumed decompositions are structured such that they can be modified via an estimator. In the estimator, the initial input matrix and plant are updated via their respective correction term $\{L_1, L_2\}$:

$$L(t) = \Delta L + L_* \xrightarrow[t\to\infty]{} L(t) = L_*, \tag{6}$$

where ΔL is the variability-uncertainty term. If both variability term converges to zero, $\{\Delta L_1, \Delta L_2\} \xrightarrow[t\to\infty]{} 0$, then the input matrix and true plant (or energy equivalent) have been numerically captured. For the control scheme to apply, the true and reference systems must be ASD and SPR, respectively.

C. Estimated State Error

Given that the true plant (A) and input matrix (B) experiences a health change caused by age or use, and the internal state (x) is often blended into a linear combination or missing from the external state (y), an estimator can be created using the reference model:

$$\text{Estimator} \begin{cases} \dot{\hat{x}} = A_m\hat{x} + B_m(L_1u + L_2y) \\ \hat{y} = C\hat{x}. \end{cases} \tag{7}$$

To minimize the error between the true and estimated systems, consider the following error state equations:

$$\begin{cases} e_x = \hat{x} - x \\ \hat{e}_y = \hat{y} - y = Ce_x. \end{cases} \tag{8}$$

To capture the internal state of the true system, the internal state error must converge to zero as time approaches infinity. To investigate the internal state error dynamics, take the time derivative of the internal state error and substitute Eq. (2) and Eq. (7):

$$\begin{aligned} \dot{e}_x &= \dot{\hat{x}} - \dot{x} \\ &= A_m\hat{x} + B_m(L_1u + L_2y) - (Ax + Bu). \end{aligned} \tag{9}$$

From Eq. (9), consider the difference between input matrices:

$$\begin{aligned} B_m(\underbrace{\Delta L_1 + L_{1*}}_{=L_1})u - \underbrace{B_m L_{1*}}_{=B} u &= B_m \underbrace{\Delta L_1 u}_{=w_u} \\ &= B_m w_u. \end{aligned} \tag{10}$$

Again, using Eq. (9) as a reference, consider the difference between the model and true plants, where $A \equiv A_m + B_m L_{2*} C$:

$$\begin{aligned} A_m\hat{x} + B_m(\underbrace{\Delta L_2 + L_{2*}}_{=L_2})y - Ax &= A_m e_x + B_m \underbrace{\Delta L_2 y}_{=w_y} \\ &= A_m e_x + B_m w_y. \end{aligned} \tag{11}$$

Therefore, the error system can be written as:

$$\begin{cases} \dot{e}_x = A_m e_x + B_m(w_u + w_y) \\ \hat{e}_y = Ce_x. \end{cases} \tag{12}$$

Additionally, the estimator can be extended to use a fixed gain (K):

$$\begin{cases} \dot{\hat{x}} = A_m\hat{x} + B_m(L_1u + L_2y) + K(y - \hat{y}) \\ \hat{y} = C\hat{x}. \end{cases} \tag{13}$$

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Resulting in the following error equation:

$$\begin{cases} \dot{e}_x = (\underbrace{A_m - KC}_{=A_c})e_x + B_m(w_u + w_y) \\ \hat{e}_y = Ce_x. \end{cases} \tag{14}$$

To use Eq. (14), find a fixed gain $(K) \ni \text{Re}(\sigma\{A_m - KC\}) < 0$.

Regardless of the estimator selected, the internal state error (e_x) can not be guaranteed to converge such that $e_x \xrightarrow[t\to\infty]{} 0$ due to the residual terms $\{w_u, w_y\}$ existing in the error equation. To adequately address these residual components, additional considerations are needed.

D. Lyapunov Stability for the Estimated State Error

Lyapunov stability analysis represents dynamic systems in terms of energy-like functions to describe the convergence of a particular or a set of states. For this case study, Lyapunov stability is used to guarantee the convergence of internal state error $(e_x) \ni e_x \xrightarrow[t\to\infty]{} 0$.

Given the state error equation as described in Eq. (12), consider the following energy-like Lyapunov equation with assumed real scalars:

$$V_e = \frac{1}{2}e_x^\dagger P_x e_x; P_x > 0, \tag{15}$$

where the $(\cdot)^\dagger$ is the conjugate transpose and where $P_x > 0$ represents a matrix P_x that is symmetric $(P_x = P_x^\dagger)$ and positive-definite $(\text{Re}(\sigma\{P_x\}) > 0)$.

To determine the energy-like time rate of change of V_e, take the time derivative of V_e and substitute Eq. (12) for the error dynamics:

$$\begin{aligned} 2\dot{V}_e =& \dot{e}_x^\dagger P_x e_x + e_x^\dagger P_x \dot{e}_x \\ =& (A_m e_x + B_m(w_u + w_y))^\dagger P_x e \\ &+ e_x^\dagger P_x (A_m e_x + B_m(w_u + w_y)) \\ =& e_x^\dagger (A_m^\dagger P_x + A_m P_x) e_x + 2\underbrace{e_x^\dagger P_x B_m(w_u + w_y)}_{=(B_m(w_u+w_y))^\dagger P_x e_x}. \end{aligned} \tag{16}$$

Modifying the SPR stability condition for the reference model:

$$\begin{cases} A_m^\dagger P_x + P_x A_m = -Q_x \\ P_x B_m = C^\dagger \end{cases} ; Q_x > 0. \tag{17}$$

From here, the SPR condition can be applied to Eq. (16), resulting in:

$$\begin{aligned} 2\dot{V}_e &= -e_x^\dagger Q_x e_x + 2\underbrace{e_x^\dagger C^\dagger}_{=\hat{e}_y^\dagger}(w_u + w_y) \\ &= -e_x^\dagger Q_x e_x + 2\hat{e}_y^\dagger w_u + 2\hat{e}_y^\dagger w_y \\ &= -e_x^\dagger Q_x e_x + 2\underbrace{(\hat{e}_y, w_u)}_{=(w_u,\hat{e}_y)} + 2\underbrace{(\hat{e}_y, w_y)}_{=(w_y,\hat{e}_y)}. \end{aligned} \tag{18}$$

By removing the residual terms $\{(\hat{e}_y.w_u), (\hat{e}_y, w_y)\}$ in Eq. (18), results in $\dot{V}_e \leq 0$.

To counter the residual terms, consider the following energy-like functions:

$$V_u + V_y = \frac{1}{2}\text{tr}(\Delta L_1 \gamma_u^{-1} \Delta L_1^\dagger) + \frac{1}{2}\text{tr}(\Delta L_2 \gamma_y^{-1} \Delta L_2^\dagger), \tag{19}$$

where $\{\gamma_u, \gamma_y\} > 0$. The energy-like time rate of change for $V_u + V_y$ follows:

$$\dot{V}_u + \dot{V}_y = \underbrace{\text{tr}(\Delta\dot{L}_1 \gamma_u^{-1} \Delta L_1^\dagger)}_{=\text{tr}(\Delta L_1 \gamma_u^{-1} \Delta\dot{L}_1^\dagger)} + \underbrace{\text{tr}(\Delta\dot{L}_2 \gamma_y^{-1} \Delta L_2^\dagger)}_{=\text{tr}(\Delta L_2 \gamma_y^{-1} \Delta\dot{L}_2^\dagger)}. \tag{20}$$

A control law for the input matrix and plant variance time rate of change $\{\Delta\dot{L}_1, \Delta\dot{L}_2\}$ can be defined as the following:

$$\begin{cases} \Delta\dot{L}_1 = -\hat{e}_y u^\dagger \gamma_u \\ \Delta\dot{L}_2 = -\hat{e}_y y^\dagger \gamma_y. \end{cases} \tag{21}$$

Substituting Eq. (21) into Eq. (20):

$$\begin{aligned} \dot{V}_u + \dot{V}_y =& \text{tr}(\underbrace{-\hat{e}_y u^\dagger \gamma_u}_{\Delta\dot{L}_1} \gamma_u^{-1} \Delta L_1^\dagger) \\ &+ \text{tr}(\underbrace{-\hat{e}_y y^\dagger \gamma_y}_{\Delta\dot{L}_2} \gamma_y^{-1} \Delta L_2^\dagger) \\ =& -\text{tr}(\hat{e}_y \underbrace{u^\dagger \Delta L_1^\dagger}_{=w_u^\dagger}) - \text{tr}(\hat{e}_y \underbrace{y^\dagger \Delta L_2^\dagger}_{=w_y^\dagger}) \\ =& -\text{tr}(\hat{e}_y w_u^\dagger) - \text{tr}(\hat{e}_y w_y^\dagger) \\ =& -\text{tr}(w_u^\dagger \hat{e}_y) - \text{tr}(w_y^\dagger \hat{e}_y) \\ =& -w_u^\dagger \hat{e}_y - w_y^\dagger \hat{e}_y \\ =& -(w_u, \hat{e}_y) - (w_y, \hat{e}_y). \end{aligned} \tag{22}$$

For notation purposes, allow the following:

$$\begin{cases} V_{euy} = V_e + V_u + V_y \\ \dot{V}_{euy} = \dot{V}_e + \dot{V}_u + \dot{V}_y. \end{cases} \tag{23}$$

From here, the estimate state error closed-loop energy-like function can be written as:

$$\begin{aligned} V_{euy} =& \frac{1}{2}e_x^\dagger P_x e_x + \frac{1}{2}\text{tr}(\Delta L_1 \gamma_u^{-1} \Delta L_1^\dagger) \\ &+ \frac{1}{2}\text{tr}(\Delta L_2 \gamma_y^{-1} \Delta L_2^\dagger). \end{aligned} \tag{24}$$

Therefore, the estimated state error closed-loop energy-like time rate of change can be written as:

$$\begin{aligned} \dot{V}_{euy} =& -\frac{1}{2}e_x^\dagger Q_x e_x + (w_u, \hat{e}_y) + (w_y, \hat{e}_y) \\ &- (w_u, \hat{e}_y) - (w_y, \hat{e}_y) \\ =& -\frac{1}{2}e_x^\dagger Q_x e_x \leq 0. \end{aligned} \tag{25}$$

Having $\dot{V}_{euy} \leq 0$ means that $\{e_x, \Delta L_1, \Delta L_2\}$ are guaranteed to be bounded. Due to $\dot{V}_{euy}$ negative-semi-definite nature, no additional information can be said about the error internal state (e_x) converging $\ni e_x \xrightarrow[t\to\infty]{} 0$.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

E. Applying Barbalat-Lyapunov Lemma on $\dot{V}_{euy}$

To guarantee $e_x \xrightarrow[t\to\infty]{} 0$, consider Barbalat-Lyapunov Lemma - Given:

1) V is lower bounded.
2) $\dot{V}$ is negative-semi-definite.
3) $\dot{V}$ is uniformly continuous in time.

If all conditions are met, then $\dot{V} \xrightarrow[t\to\infty]{} 0$ according to [10].

The first two conditions of the Barbalat-Lyapunov Lemma are satisfied with Eq. (24) and Eq. (25). The third condition, $\dot{V}_{euy}$ being uniformly continuous in time, can be satisfied by showing that $\ddot{V}_{euy}$ is bounded [10].

To prove $\ddot{V}_{euy}$ is bounded, consider W_{euy}:

$$W_{euy} \geq -2\dot{V}_{euy} \geq 0. \tag{26}$$

Taking the time derivative of W_{euy} results in the following:

$$\begin{aligned}\dot{W}_{euy} =& 2e_x^\dagger Q_x \dot{e}_x \\ =& 2e_x^\dagger Q_x (A_m e_x + B_m(w_u + w_y)) \\ =& 2e_x^\dagger Q_x (A_m e_x + B_m(\Delta L_1 u + \Delta L_2 y)).\end{aligned} \tag{27}$$

From Eq. (25), $\{e_x, \Delta L_1, \Delta L_2\}$ are bounded. Input (u) can be any bounded-continuous waveform. Following, the true plant is assumed stable (i.e., $\mathrm{Re}(\sigma\{A\}) < 0$); therefore, a bounded input will result in a bounded output (y) [11]. Combining all bounded results yields: $\dot{W}_{euy}$ is indeed bounded. Making $\ddot{V}_{euy}$ bounded.

Given that all the conditions of Barbalat-Lyapunov are satisfied, $\dot{V}_{euy}$ evolution in time can be expressed as:

$$\dot{V}_{euy} \xrightarrow[t\to\infty]{} 0. \tag{28}$$

Therefore, proves $e_x \xrightarrow[t\to\infty]{} 0$ is asymptotically guaranteed. However, regardless of Barbalat-Lyapunov being satisfied, Lyapunov stability results only guarantees $\{\Delta L_1, \Delta L_2\}$ to be bounded. If $\{\Delta L_1, \Delta L_2\} \xrightarrow[t\to\infty]{} 0$ numerically, the true input matrix and plant or an energy equivalence have been captured. Additionally, without loss of generality, the derived Lyapunov stability proof can be modified for the error system using fixed gain, Eq. (12).

Altogether, assuming the reference (A_m, B_m, C) and true (A, B, C) systems are SPR and ASD respectfully, such that the decomposition of the true input matrix (B) and plant (A) can be written as $B \equiv B_m L_{1*}$ and $A \equiv A_m + B_m L_{2*} C$. Then adaptive laws (Eq. (21)) and diagram (Fig. 1) can be formulated such that the internal state error is guaranteed to converge to zero asymptotically. Lyapunov stability proof only guarantees that $\{\Delta L_1, \Delta L_2\}$ will be bounded. However, if $\{\Delta L_1, \Delta L_2\} \xrightarrow[t\to\infty]{} 0$, then the true input matrix and plant or energy equivalent have been numerically captured.

IV. Main Result - Theorem 2

An initial attempt was made to incorporate feedback filters into Theorem 1, similar to those as described in [2]. However, numerical results were not fruitful. See Section VII for additional details. This section further extends *Griffith* 2022 findings by incorporating the feedback filter into the input matrix estimation law [8]. The addition of the feedback filter to the input matrix estimation law enables faster convergence of the internal state error (e_x) compared to the non-feedback filter case numerically. The tradeoff of using the feedback filter is that the internal state (e_x) can only be guaranteed to converge to a neighborhood about zero $(e_x \xrightarrow[t\to\infty]{} R_*)$, versus the internal error converging to zero $(e_x \xrightarrow[t\to\infty]{} 0)$ in the non-feedback filter case.

A key difference between Theorem 1 and 2 is knowledge of the true system's plant (A). To apply Theorem 2, the true system plant (A) must be known, while the input matrix model (B_m) must remain. The introduction of the feedback term in the input matrix estimator law disables the use of the Barbalat-Lyapunov Lemma and increases the complexity of the analysis. Much of the analysis consists of bounding terms using a combination of algebraic manipulation, Sylvester's Inequality, and Cauchy-Schwarz inequality. Applying the feedback filter to the adaptive input matrix estimator is summarized in Theorem 2 and Fig. 2.

Theorem 2: Output Feedback on Reference Model for Adaptive Input Matrix using a Feedback Filter and State Estimation. Consider the following state error system:

$$\begin{cases} \dot{e}_x = (A - KC)e_x + B_m \Delta L_1 u \\ \hat{e}_y = Ce_x = C(\hat{x} - x) \\ L_1 = \Delta L_1 + L_{1*} \\ \dot{L}_1 = \Delta \dot{L}_1 = -\hat{e}_y u^\dagger \gamma_u - \alpha L_1 \end{cases}, \tag{29}$$

where e_x is the estimated internal state error, $\hat{e}_y$ is the external estimated state error, L_{1*} is fixed-correction matrix, ΔL_1 is the variability-uncertainty terms, K is a fixed gain, and $\gamma_u > 0$ is the interaction tuning term. Given:

1) The triples of (A, B, C) and (A, B_m, C) are both SPR.
2) A true plant (A) must exist.
3) A model input matrix (B_m) must exist.
4) Output matrix (C) is known.
5) Allow $B \in \mathrm{Sp}\{B_m L_{1*}\} \ni B \equiv B_m L_{1*}$.
6) The set of eigenvalues (σ) of the true plant are stable (i.e. $\mathrm{Re}(\sigma(A)) < 0$).

If conditions are met, then the internal error state converges to a neighborhood about zero $(e_x \xrightarrow[t\to\infty]{} R_*)$ asymptotically. ΔL_1 is guaranteed to be bounded; however, no guarantee of $\Delta L_1 \xrightarrow[t\to\infty]{} 0$. If $\Delta L_1 \xrightarrow[t\to\infty]{} 0$, then the true system input matrix or some energy equivalence has been numerically captured.

A. Defining the True and Reference System

The true system dynamics will follow all of the assumptions and constraints outlined in Section III-A. In contrast, allow the reference model to have the following state space characteristics:

$$\text{Reference Model} \begin{cases} \dot{x}_m = Ax_m + B_m u \\ y_m = Cx_m, \end{cases} \tag{30}$$

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

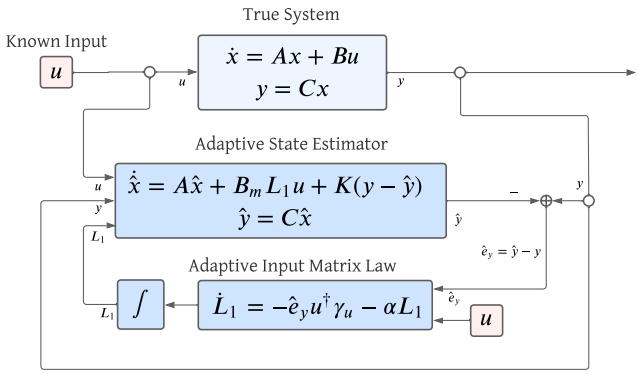


Figure 2. Control diagram for adaptive input matrix with a feedback filter and state estimation given a known input (u), output matrix (C), and external state (y).

where the true plant (A) is assumed to be known. The input matrix model (B_m) can be corrected using a fixed correction term as outlined in Eq. (4) and Eq. (6). For the control scheme to apply, the true and reference systems must be SPR.

B. Estimated State Error

Assuming that the input matrix (B) only experiences a form of health change and the true internal states (x) are often blended into some weighted linear combination of the external state (y), an estimator can be formed using the reference model:

$$\text{Estimator} \begin{cases} \dot{\hat{x}} = A\hat{x} + B_m L_1 u \\ \hat{y} = C\hat{x}. \end{cases} \tag{31}$$

Following a similar protocol as described in Section III-C, error states are defined (Eq. (8)) and the error system dynamics follow:

$$\begin{cases} \dot{e}_x = Ae_x + B_m \underbrace{(\Delta L_1 u)}_{=w_u} \\ \hat{e}_y = Ce_x. \end{cases} \tag{32}$$

The estimator defined in Eq. (31) can be extended to use a fixed gain (K):

$$\begin{cases} \dot{\hat{x}} = A\hat{x} + B_m L_1 u + K(y - \hat{y}) \\ \hat{y} = C\hat{x}. \end{cases} \tag{33}$$

Resulting in the following error equation:

$$\begin{cases} \dot{e}_x = (\underbrace{A - KC}_{=A_c})e_x + B_m w_u \\ \hat{e}_y = Ce_x. \end{cases} \tag{34}$$

Again, regardless of the estimator selected, the estimated states cannot be guaranteed to converge to the true states due to a residual term (w_u) existing in the error equation. To accommodate the residual term, additional constraints are needed.

C. Lyapunov Stability for the Estimated States

The initial Lyapunov Stability analysis will follow a similar protocol as outlined in Section III-D. Using the same Lyapunov function as shown in Eq. (15), the energy-like time rate of change of V_e follows:

$$\begin{aligned} 2\dot{V}_e =& \dot{e}_x^\dagger P_x e_x + e_x^\dagger P_x \dot{e}_x \\ =& (Ae_x + B_m w_u)^\dagger P_x e + e_x^\dagger P_x (Ae_x + B_m w_u) \\ =& e_x^\dagger (A^\dagger P_x + AP_x) e_x + 2 \underbrace{e_x^\dagger P_x B_m w_u}_{=(B_m w_u)^\dagger P_x e_x}. \end{aligned} \tag{35}$$

Modifying the SPR stability condition for the reference model:

$$\begin{cases} A^\dagger P_x + P_x A = -Q_x \\ P_x B_m = C^\dagger \end{cases} ; Q_x > 0. \tag{36}$$

Now, the SPR condition can be applied onto Eq. (35):

$$\begin{aligned} 2\dot{V}_e &= -e_x^\dagger Q_x e_x + 2 \underbrace{e_x^\dagger C^\dagger}_{=\hat{e}_y^\dagger} w_u \\ &= -e_x^\dagger Q_x e_x + 2\hat{e}_y^\dagger w_u \\ &= -e_x^\dagger Q_x e_x + 2 \underbrace{(\hat{e}_y, w_u)}_{=(w_u, \hat{e}_y)}. \end{aligned} \tag{37}$$

By removing the residual term $((w_u, \hat{e}_y))$ in Eq. 37, results in $\dot{V}_e \leq 0$.

To counter the residual term, consider the following energy-like function:

$$V_u = \frac{1}{2}\text{tr}(\Delta L_1 \gamma_u^{-1} \Delta L_1^\dagger), \tag{38}$$

where $\gamma_u > 0$. The energy-like time rate of change for V_u follows:

$$\dot{V}_u = \underbrace{\text{tr}(\Delta \dot{L}_1 \gamma_u^{-1} \Delta L_1^\dagger)}_{=\text{tr}(\Delta L_1 \gamma_u^{-1} \Delta \dot{L}_1^\dagger)}. \tag{39}$$

A control law for the input matrix variance time rate of change $(\Delta \dot{L}_1)$ can be defined as:

$$\Delta \dot{L}_1 = -\hat{e}_y u^\dagger \gamma_u - \alpha L_1, \tag{40}$$

where αL_1 acts as a feedback filter. Additional constraints will be added for α later in the derivation, but for now, allow α to be a scalar with a value greater than zero $(\alpha > 0)$.

Substituting Eq. (40) into Eq. (39):

$$\begin{aligned} \dot{V}_u &= \text{tr}(\underbrace{(-\hat{e}_y u^\dagger \gamma_u - \alpha L_1)}_{\Delta \dot{L}_1} \gamma_u^{-1} \Delta L_1^\dagger) \\ &= \text{tr}(-\hat{e}_y u^\dagger \gamma_u \gamma_u^{-1} \Delta L_1^\dagger - \alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger) \\ &= -\text{tr}(\hat{e}_y \underbrace{u^\dagger \Delta L_1^\dagger}_{=w_u^\dagger}) - \text{tr}(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger) \\ &= -\text{tr}(\underbrace{\hat{e}_y w_u^\dagger}_{=w_u^\dagger \hat{e}_y}) - \text{tr}(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger) \\ &= -(w_u, \hat{e}_y) - \text{tr}(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger). \end{aligned} \tag{41}$$

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Therefore, the closed-loop energy-like function can be written as:

$$V_{eu} = V_e + V_u = \frac{1}{2}e_x^\dagger P_x e_x + \frac{1}{2}\text{tr}(\Delta L_1 \gamma_u^{-1} \Delta L_1^\dagger). \quad (42)$$

While the closed-loop energy-like time rate of change follows:

$$\begin{aligned}\dot{V}_{eu} = \dot{V}_e + \dot{V}_u =& -\frac{1}{2}e_x^\dagger Q_x e_x + (w_u, \hat{e}_y) - (w_u, \hat{e}_y) \\ &- \text{tr}(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger) \\ =& -\frac{1}{2}e_x^\dagger Q_x e_x - \text{tr}(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger).\end{aligned} \quad (43)$$

Due to a residual term existing Eq. (43) ($-\text{tr}(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger)$) Barabalat-Lyapunov methods, as implemented in Section III-E, cannot be utilized. Therefore, $e_x \xrightarrow[t\to\infty]{} 0$ cannot be guaranteed. However, an upper limit or radius of convergence (R_*) can be determined for the internal state error (e_x).

D. Determining the Radius of Convergence for the Internal Error

The derivation to determine the radius of convergence (R_*) for the internal error (e_x) follows a similar approach as described in the *Fuentes* 2025 [2]. The approach requires bounding each component of the right-hand side of Eq. (43), then utilizing algebraic manipulation to find the limit-supremum of the internal state error as time approaches infinity.

1) Bounding $\frac{1}{2}e_x^\dagger Q_x e_x$: Applying Sylvester's Inequality to Eq. (15) results in:

$$\frac{\lambda_{\min}(P_x)}{2}\|e_x\|^2 \le V_e = \frac{1}{2}e_x^\dagger P_x e_x \le \frac{\lambda_{\max}(P_x)}{2}\|e_x\|^2, \quad (44)$$

where $\{\lambda_{\min}(P_x), \lambda_{\max}(P_x)\}$ are the minimum and maximum eigenvalues of matrix P_x. Following, $\|e_x\|$ is the norm of the internal error. The upper bound of V_e is representative of the right-hand side of Eq. (44). Therefore, the upper bound of $\|e_x\|^2$ can be expressed as:

$$\frac{2V_e}{\lambda_{\max}(P_x)} \le \|e_x\|^2. \quad (45)$$

Applying Sylvester's Inequality to $-\frac{1}{2}e_x^\dagger Q_x e_x$, the upper bound can be determined as:

$$-\frac{1}{2}e_x^\dagger Q_x e_x \le -\frac{\lambda_{\min}(Q_x)}{2}\|e_x\|^2. \quad (46)$$

Substituting Eq. (45) into Eq. (46):

$$\begin{aligned}-\frac{1}{2}e_x^\dagger Q_x e_x &\le -\frac{\lambda_{\min}(Q_x)}{2}\underbrace{\frac{2V_e}{\lambda_{\max}(P_x)}}_{=\|e_x\|^2} \\ &\le -\frac{\lambda_{\min}(Q_x)}{\lambda_{\max}(P_x)}V_e = -2\alpha V_e.\end{aligned} \quad (47)$$

Due to both $\{P_x, Q_x\} > 0$, means both have sets of positive eigenvalues and therefore require $2\alpha \ge \frac{\lambda_{\min}(Q_x)}{\lambda_{\max}(P_x)} > 0$ and be of scalar value.

2) Bounding $-tr(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger)$: In order to bound $-\text{tr}(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger)$ from Eq. (43), recall that $L_1 = \Delta L_1 + L_{1*}$. Therefore, the residual term can be updated as:

$$\begin{aligned}\text{tr}(\alpha L_1 \gamma_u^{-1} \Delta L_1^\dagger) =& \alpha \text{tr}(\underbrace{(\Delta L_1 + L_{1*})}_{=L_1} \gamma_u^{-1} \Delta L_1^\dagger) \\ =& \alpha \underbrace{\text{tr}(\Delta L_1 \gamma_u^{-1} \Delta L_1^\dagger)}_{=2V_u} \\ &+ \alpha \text{tr}(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger) \\ =& 2\alpha V_u + \alpha \text{tr}(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger).\end{aligned} \quad (48)$$

Substituting both Eq. (47) and Eq. (48) into Eq. (43) results in:

$$\dot{V}_{eu} = \underbrace{-2\alpha V_e - 2\alpha V_u}_{=-2\alpha V_{eu}} - \alpha \text{tr}(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger). \quad (49)$$

Therefore, Eq. (49) can alternatively be written as:

$$\dot{V}_{eu} + 2\alpha V_{eu} = -\alpha \text{tr}(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger). \quad (50)$$

The right-hand side of Eq. (50) can be upper bounded by taking the absolute value of terms:

$$\dot{V}_{eu} + 2\alpha V_{eu} \le \alpha |\text{tr}(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger)|. \quad (51)$$

3) Upper Bounding $|tr(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger)|$: Applying the Cauchy-Schwarz inequality on $|\text{tr}(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger)|$:

$$\begin{aligned}&|\text{tr}(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger)| \le \\ &\text{tr}(L_{1*} \gamma_u^{-1} L_{1*}^\dagger)^{\frac{1}{2}} \underbrace{\text{tr}(\Delta L_1 \gamma_u^{-1} \Delta L_1^\dagger)^{\frac{1}{2}}}_{=(2V_u)^{\frac{1}{2}} \le (2V_{eu})^{\frac{1}{2}}}.\end{aligned} \quad (52)$$

Using the trace properties and Cauchy-Schwarz inequality, $\text{tr}(L_{1*} \gamma_u^{-1} L_{1*}^\dagger)^{\frac{1}{2}}$ can be alternatively written as:

$$\begin{aligned}\text{tr}(L_{1*} \gamma_u^{-1} L_{1*}^\dagger)^{\frac{1}{2}} &\le |\text{tr}(L_{1*}^\dagger L_{1*} \gamma_u^{-1})^{\frac{1}{2}}| \\ &\le \text{tr}(L_{1*}^\dagger L_{1*})^{\frac{1}{2}} \text{tr}(\gamma_u^{-1})^{\frac{1}{2}}.\end{aligned} \quad (53)$$

For notation purposes, allow the following assumptions:

$$M_k > 0 \ni \text{tr}(L_{1*}^\dagger L_{1*})^{\frac{1}{2}} \le M_k \quad (54)$$

and

$$\text{tr}(\gamma_u^{-1}) \le \left(\frac{1}{\alpha M_k}\right)^2. \quad (55)$$

Substituting Eq. (54) and Eq. (55) into Eq. (53) results in:

$$\begin{aligned}\text{tr}(L_{1*} \gamma_u^{-1} L_{1*}^\dagger)^{\frac{1}{2}} &\le \text{tr}(L_{1*}^\dagger L_{1*})^{\frac{1}{2}} \text{tr}(\gamma_u^{-1})^{\frac{1}{2}} \\ &\le M_k \left(\frac{1}{\alpha M_k}\right) \\ &\le \frac{1}{\alpha}.\end{aligned} \quad (56)$$

Thus, substituting Eq. (56) into Eq. (52) results in:

$$\begin{aligned}|\text{tr}(L_{1*} \gamma_u^{-1} \Delta L_1^\dagger)| &\le \underbrace{\text{tr}(L_{1*} \gamma_u^{-1} L_{1*}^\dagger)^{\frac{1}{2}}}_{=\frac{1}{\alpha}} (2V_{eu})^{\frac{1}{2}} \\ &\le \frac{1}{\alpha}(2V_{eu})^{\frac{1}{2}}.\end{aligned} \quad (57)$$

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

4) Combing Bounds to Determine Radius of Convergence (R_*): Combining the results in Eq. (57) into Eq. (51) results in:

$$\begin{aligned}\dot{V}_{eu} + 2\alpha V_{eu} &\leq \alpha \underbrace{|\mathrm{tr}(L_{1*}\gamma_u^{-1}\Delta L_1^\dagger)|}_{=\frac{1}{a}(2V_{eu})^{\frac{1}{2}}} \\ &\leq (2V_{eu})^{\frac{1}{2}}.\end{aligned} \tag{58}$$

Thus, Eq. (58) can be rewritten as:

$$\frac{\dot{V}_{eu} + 2\alpha V_{eu}}{V_{eu}^{\frac{1}{2}}} \leq \sqrt{2}. \tag{59}$$

Consider the following expression:

$$\begin{aligned}\frac{d}{dt}(2e^{\alpha t}V_{eu}^{\frac{1}{2}}) &= 2(\alpha e^{\alpha t}V_{eu}^{\frac{1}{2}} + e^{\alpha t}\frac{1}{2}V_{eu}^{-\frac{1}{2}}\dot{V}_{eu}) \\ &= \frac{2\alpha e^{\alpha t}V_{eu} + e^{\alpha t}\dot{V}_{eu}}{V_{eu}^{\frac{1}{2}}} \\ &= \frac{\dot{V}_{eu} + 2\alpha V_{eu}}{V_{eu}^{\frac{1}{2}}}e^{\alpha t}.\end{aligned} \tag{60}$$

Therefore, Eq. (59) can be alternatively expressed as:

$$\frac{d}{dt}(2e^{\alpha t}V_{eu}^{\frac{1}{2}}) \leq \sqrt{2}e^{\alpha t}. \tag{61}$$

Integrating and evaluating both sides of Eq. (61) from 0 to τ results in:

$$2e^{\alpha\tau}V_{eu}(\tau)^{\frac{1}{2}} - 2e^{\alpha 0}V_{eu}(0)^{\frac{1}{2}} \leq \frac{\sqrt{2}}{\alpha}\left(e^{\alpha\tau} - e^{\alpha 0}\right). \tag{62}$$

Solving for $V_{eu}(\tau)^{\frac{1}{2}}$:

$$V_{eu}(\tau)^{\frac{1}{2}} \leq e^{-\alpha\tau}V_{eu}(0) + \frac{\sqrt{2}}{2\alpha}(1 - e^{-\alpha\tau}). \tag{63}$$

From Eq. (63), given that $\alpha > 0$, then it can be shown that $V_{eu}(\tau)^{\frac{1}{2}}$ is bounded for all $\tau \geq 0$ and therefore bounding $V_{eu}(\tau)$ for all $\tau \geq 0$. By $V_{eu}(\tau)$ being bounded for all $\tau \geq 0$, this also implies that both $\{e_x, \Delta L_1\}$ remain bounded as they compose V_{eu}.

$V_{eu}^{\frac{1}{2}}$ can be further lower bounded by considering the following inequality:

$$\frac{\lambda_{\min}(P_x)^{\frac{1}{2}}}{\sqrt{2}}\|e_x\| \leq V_e^{\frac{1}{2}} + V_u^{\frac{1}{2}} \leq V_{eu}(\tau)^{\frac{1}{2}}. \tag{64}$$

Solving for $\|e_x\|$ in Eq. (64) results in:

$$\begin{aligned}\|e_x\| \leq &\frac{\sqrt{2}}{\lambda_{\min}(P_x)^{\frac{1}{2}}}e^{-a\tau}V_{eu}(0) \\ &+ \frac{1}{\alpha\lambda_{\min}(P_x)^{\frac{1}{2}}}(1 - e^{-\alpha\tau}).\end{aligned} \tag{65}$$

Taking the limit and supremum (upper bound) of $\|e_x\|$ in Eq. (65) results in the radius of confidence (R_*) as the adaptive input matrix estimator with feedback term is used:

$$\limsup_{\tau\to\infty}\|e_x\| = \frac{1}{\alpha\lambda_{\min}(P_x)^{\frac{1}{2}}} \equiv R_*. \tag{66}$$

Recall $\|e_x\|$ being bounded is based on the assumption of both Eq. (54) and Eq. (55). Equation (55) can notably be upper bounded by applying Sylvester's Inequality, altering the radius of confidence by a scaling factor.

Following, V_u can be bounded by applying the properties of the trace and the Cauchy–Schwarz inequality:

$$V_u = \frac{1}{2}\mathrm{tr}(\Delta L_1^\dagger\Delta L_1\gamma_u^{-1}) \leq \frac{1}{2}\underbrace{\mathrm{tr}(\Delta L_1^\dagger\Delta L_1)}_{=\|\Delta L_1\|_F^2}\mathrm{tr}(\gamma_u^{-1}), \tag{67}$$

where $\|\Delta L_1\|_F$ is the Frobenius norm of ΔL_1. Using the assumption of Eq. (55), $V_u^{\frac{1}{2}}$ can be expressed as:

$$V_u^{\frac{1}{2}} = \frac{1}{\sqrt{2}}\|\Delta L_1\|_F\left(\frac{1}{\alpha M_k}\right). \tag{68}$$

Using a similar procedure for deriving the limit and supremum of $\|e_x\|$, this can be done to determine the limit and supremum of $\|\Delta L_1\|_F$:

$$\frac{1}{\sqrt{2}}\|\Delta L_1\|_F\left(\frac{1}{\alpha M_k}\right) \leq V_e^{\frac{1}{2}} + V_u^{\frac{1}{2}} \leq V_{eu}(\tau)^{\frac{1}{2}}. \tag{69}$$

Resulting in the following radius of confidence:

$$\limsup_{\tau\to\infty}\|\Delta L_1\| = M_k \equiv R_{**}. \tag{70}$$

E. Summarizing Results

Assuming that the true (A, B, C) and reference (A, B_m, C) systems are SPR and the true input matrix satisfies $B \equiv B_m L_{1*}$, then the adaptive input matrix estimation law using the feedback filter (Eq. (40)) and diagram (Fig. 2) guarantees $\|e_x\| \xrightarrow[t\to\infty]{} R_*$ asymptotically. Following, Lyapunov Stability Analysis only guarantees that ΔL_1 will remain bounded. Finally, the Lyapunov stability proof can also be applied to the error system under the use of fixed gains in Eq. (34).

V. Illustrative Example - Theorem 1

The following is an illustrative example of applying Theorem 1 and the control diagram (Fig. 1) on a general case study. Numerical values for (A_m, B_m, C) and (A, B, C) are derived and modified from [12].

A. State Space Representations for Reference and True Systems

Allow the reference model as defined in Eq. (3) have the following properties:

$$\begin{aligned}A_m &= \begin{bmatrix}-7 & 2 & 4\\ -2 & -1 & 2\\ -2 & 2 & -1\end{bmatrix}; \\ B_m &= \begin{bmatrix}0\\ .7\\ 2\end{bmatrix}; C = \begin{bmatrix}0.5 & 0 & 1\end{bmatrix}; x(0) = 0.\end{aligned} \tag{71}$$

To apply the control scheme as defined in Theorem 1 and show in Fig. 1, allow the true system as defined by Eq. (2) have the following properties:

1) $B \in \mathrm{Sp}\{B_m L_{1*}\} \ni B \equiv BL_{1*}$.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

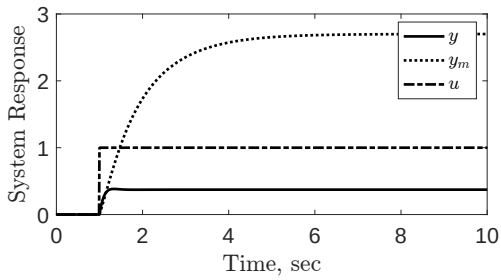


Figure 3. True (y) and reference model (y_m) output response given a unit step input (u).

2) $A \in \text{Sp}\{A_m, B_m L_{2*} C\} \ni A \equiv A_m + B L_{2*} C$.

Assume the health change for the input matrix and the plant can be described by $\{L_{1*}, L_{2*}\} \ni L_{1*} = 2$ and $L_{2*} = -5$. Therefore, the true system can be defined by the following:

$$\begin{aligned} A \equiv & A_m + B_m L_{2*} C = \begin{bmatrix} -7 & 2 & 4 \\ -3.75 & -1 & -1.5 \\ -7 & 2 & -11 \end{bmatrix}; \\ B \equiv & B_m L_{1*} = \begin{bmatrix} 0 \\ 1.4 \\ 4 \end{bmatrix}; C = \begin{bmatrix} 0.5 & 0 & 1 \end{bmatrix}; x(0) = 0. \end{aligned} \tag{72}$$

Recall that the constitutive constants of the true plant (A) and input matrix (B) are unknown. However, an initial estimate of the plant (A_m) and input matrix (B_m) exists.

When both the reference and true systems, as defined in Eq. (71) and Eq. (72), are given a unit step input, as shown in Fig. 3, the differences in rise times and output response become evident. These differences can be further explained by examining the eigenvalues of the reference and true plants:

$$\begin{aligned} \sigma(A_m) &= \{-1, -3, -5\} \\ \sigma(A) &\approx \{-2.28, -8.36 \pm i5.05\}. \end{aligned} \tag{73}$$

B. Defining the Known Input (u)

To implement the control scheme, a bounded and continuous input must be used. In practice, this input can be a known disturbance. For this example, allow the input to be defined as:

$$u = 2 + \sin(2t). \tag{74}$$

C. Adaptive State and Input Matrix Estimation

In this section, the presented control scheme, detailed in Fig. 1, is implemented with two cases: with and without the use of a fixed gain (K) term.

1) Adaptive Control Scheme without the use of Fixed Gain $(K = 0)$: The control scheme detailed in Fig. 1 is implemented without using the fixed gain term $(K = 0)$ and $\{\gamma_u, \gamma_y\} = I$. As derived in the proof, Fig. 4 demonstrates the convergence of the internal state, where $e_x \xrightarrow[t\to\infty]{} 0$. Given that the internal state error converges to zero, equivalently, the external state error converges $\ni \hat{e}_y \xrightarrow[t\to\infty]{} 0$. Meaning that the estimated output $(\hat{y})$ converges to the true output (y).

Although the proof only guarantees that the adaptive variance will be bounded, numerically $\{\Delta L_1, \Delta L_2\} \xrightarrow[t\to\infty]{} 0$. For this case study, the true input matrix and plant have been numerically captured, Fig. 5 and Fig. 6.

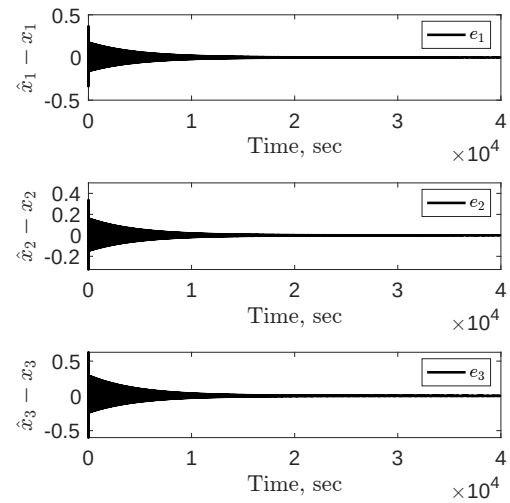


Figure 4. Internal state error converging to zero without the use of the fixed gain $(K = 0)$.

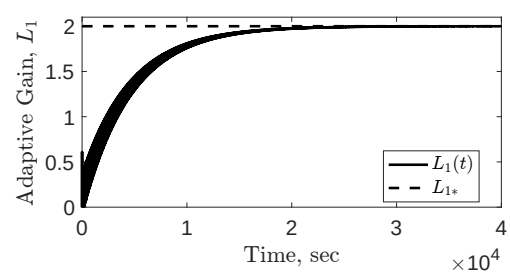


Figure 5. Input Matrix adaptive term $L_1(t)$ converging to L_{1*} without the use of the fixed gain $(K = 0)$.

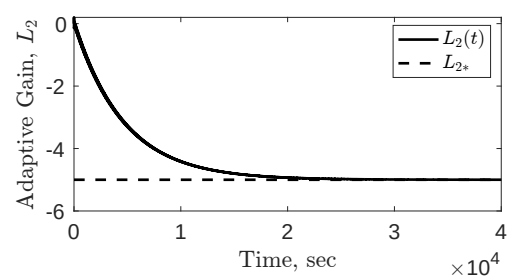


Figure 6. Plant correction adaptive term $L_2(t)$ converging to L_{2*} without the use of the fixed gain $(K = 0)$.

2) Adaptive Control Scheme with the use of Fixed Gain $(K \neq 0)$: The control scheme detailed in Fig. 1 is implemented using the fixed gain term $(K \neq 0)$ and $\{\gamma_u, \gamma_y\} = 1$. The fixed gain term K was derived using a Linear Quadratic Regulator (LQR) where $Q = I_3$ and $R = 1$. Similarly to the result of Section V-C1, $e_x \xrightarrow[t\to\infty]{} 0$, shown in Fig. 7. Again, since the internal state error converges to zero, the external

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

error will converge to zero for the true and estimator systems. Moreover, as $\{\Delta L_1, \Delta L_2\} \xrightarrow[t\to\infty]{} 0$, the true input matrix and plant are numerically captured in Fig. 8 and Fig. 9.

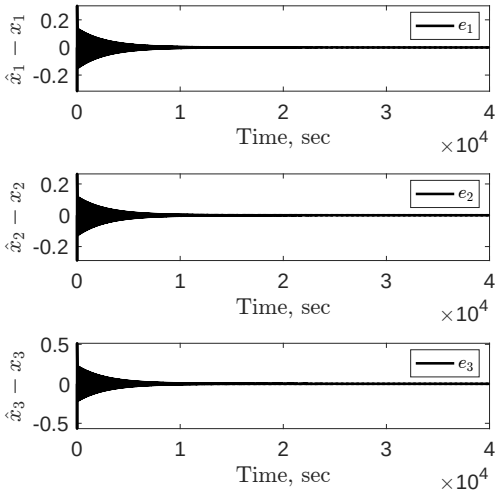


Figure 7. Internal state error converging to zero with the use of the fixed gain $(K \neq 0)$.

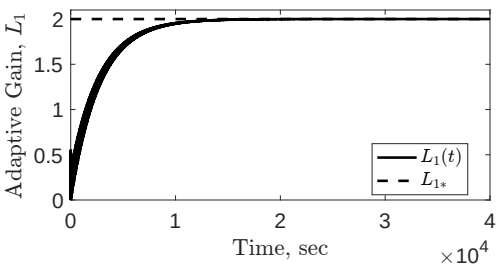


Figure 8. Input Matrix adaptive term $L_1(t)$ converging to L_{1*} with the use of the fixed gain $(K \neq 0)$.

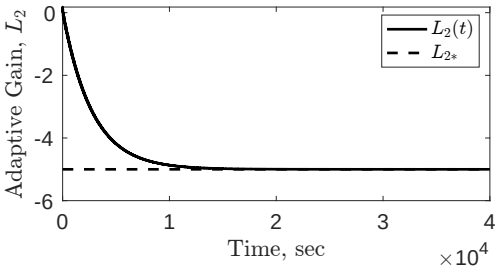


Figure 9. Plant correction adaptive term $L_2(t)$ converging to L_{2*} with the use of the fixed gain $(K \neq 0)$.

There can be a benefit of using a fixed gain term in the estimator, as the term can affect the time in which internal states and adaptive terms converge, compare Fig. 5 and Fig.8. More crucially, both adaptive tuning terms $\{\gamma_u, \gamma_y\}$ can be adjusted to amplify or dampen the effects of the adaptive controller, directly impacting the convergence of the error state. For this particular example, setting $\gamma_u = 1.3$ and $\gamma_y = 1.85$ reduces the time in which $e_x \xrightarrow[t\to\infty]{} 0$ and $L \xrightarrow[t\to\infty]{} L_*$ by order of magnitude faster than the depicted figures in this text. However, there are numerical limits for the tuning terms $\{\gamma_u, \gamma_y\}$. Making the adaptive controller too sensitive to changes may lead to divergent artifacts.

3) Kalman-Bucy Filter-Kalman Filter in Continuous Time: The previous two sections, Sections V-C1 and V-C2, illustrate an example that implements an adaptive estimator both with and without a fixed gain (K) term. When the adaptive laws are removed, and only the fixed gain estimator is used, the reference model is no longer updated within the estimator. This is evident in Fig. 10 and Fig. 11, where it can be seen that both the internal and external errors reach a non-zero steady-state value. The value of the fixed gain term (K) is derived using LQR, where $Q = I_3$ and $R = 1$. Although the observed time window for the error dynamics response is limited, the system response does not change as time approaches infinity.

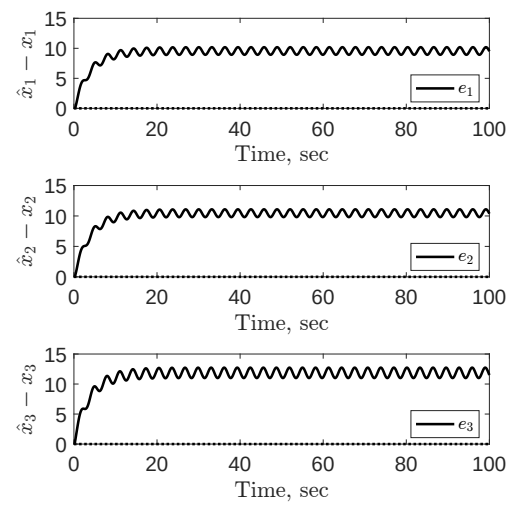


Figure 10. Internal state error (e_x) not converging in the absence of the adaptive estimator. Only the fixed gain (K) is used in the estimator.

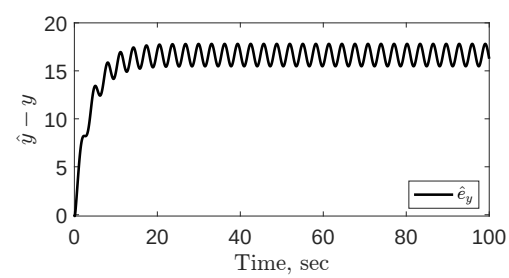


Figure 11. External state error $(\hat{e}_y)$ not converging in the absence of the adaptive estimator.

VI. Illustrative Example - Theorem 2

The following is an illustrative example of applying Theorem 2 and the control diagram (Fig. 2) on a general case study. Numerical values for (A, B_m, C), (A, B, C), and input (u) are the same from Section V.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A. Input Matrix Estimation with the Feedback Filter ($\alpha > 0$ and $K = 0$)

In this section, the control scheme outlined in Fig. 2 is implemented with only one case study, without the use of the fixed gain ($K = 0$) term. The reason for this is that Section V already considers the case where the adaptive laws are coupled with and without fixed gain terms.

As derived in the proof, Fig. 12 shows that the internal state error converges to a neighborhood of zero when the adaptive input matrix law is coupled with the feedback term ($\|e_x\| \xrightarrow[t\to\infty]{} R_*$). As the internal-state error settles within this neighborhood, the external-state error correspondingly follows. In Fig. 13, we see that L_1 approaches L_{1*}, but does not fully converge, consistent with the theoretical result, which guarantees only that the adaptive variance (ΔL_1) remains bounded. Although, there exist variance in L_1, the updated input matrix ($B_m L_1$) more closely aligns with the true input matrix ($B \equiv B_m L_{1*}$) than the nominal input matrix model (B_m). Values used on the adaptive input matrix estimation law follows: $\alpha = .05$ and $\gamma_u = I$.

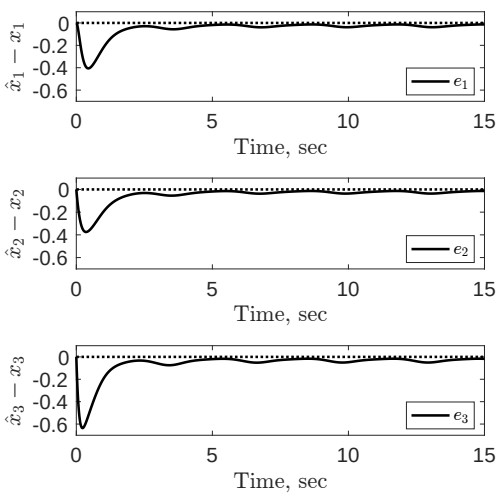


Figure 12. Internal state error converging to the neighborhood about zero without the use of the fixed gain ($K = 0$) and using the feedback filter on the adaptive input matrix law.

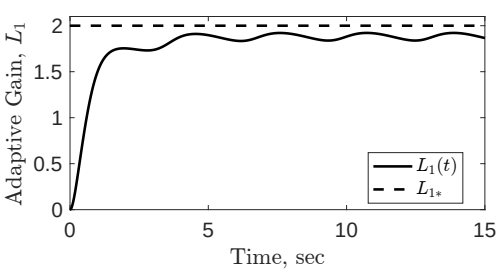


Figure 13. Input Matrix adaptive term $L_1(t)$ approaching L_{1*} without the use of the fixed gain ($K = 0$) and using the feedback filter on the adaptive input matrix law.

B. Input Matrix Estimation without the Feedback Filter ($\alpha = 0$ and $K = 0$)

The control scheme detailed in Fig. 2 is implemented without using a fixed gain term ($K = 0$) and without the feedback filter ($\alpha = 0$). After truncation, proof follows the results presented in *Griffith* 2022 [8]. The adaptive input matrix estimation law uses $\gamma_u = I$. In the absence of the feedback filter, $\|e_x\| \xrightarrow[t\to\infty]{} 0$, as demonstrated in Fig. 14, consistent with theoretical results. As the internal state error converges to zero, the external state error will follow. Lastly, as $\Delta L_1 \xrightarrow[t\to\infty]{} 0$, the true input matrix is captured, as shown in Fig. 15. Note that theoretical results only guarantee ΔL_1 to remain bounded, whereas the numerical results exhibit full convergence.

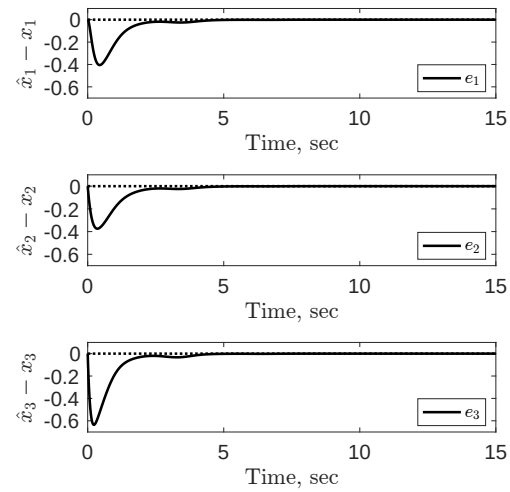


Figure 14. Internal state error converging to zero without the use of the fixed gain ($K = 0$) and not using the feedback filter on the adaptive input matrix law.

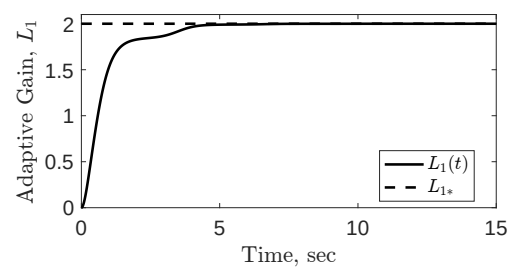


Figure 15. Input Matrix adaptive term $L_1(t)$ converging to L_{1*} without the use of the fixed gain ($K = 0$) and not using the feedback filter on the adaptive input matrix law.

C. Kalman-Bucy Filter-Kalman Filter in Continuous Time

The previous two sections, Sections VI-A and VI-B, presented examples of implementing the adaptive input-matrix estimator both with and without a feedback term, while not using a fixed gain term ($K = 0$). When the adaptive laws are removed, and only the fixed gain term is retained, the input matrix model is no longer updated within the estimator. This

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

lack of adaptation results in larger internal and external state errors compared to the adaptive cases, as shown in Fig. 16 and Fig. 17. The value for the fixed gain term is obtained using LQR, where $Q = I_3$ and $R = 1$.

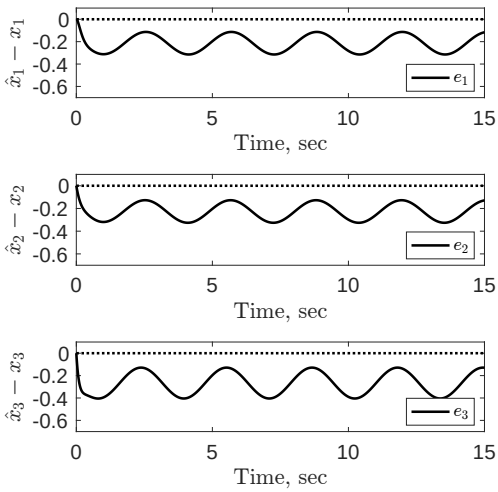


Figure 16. Internal state error (e_x) not converging to zero in the absence of the adaptive input matrix estimator. Only fixed gain (K) used in the estimator.

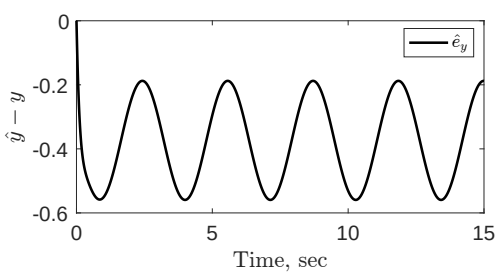


Figure 17. External state error $(\hat{e}_y)$ not converging to zero in the absence of the adaptive input matrix estimator.

VII. Discussion

The theorem presented in this text pertains to globally stable LTI systems. By utilizing local linearization, the theorems can be applied to both linear and non-linear systems that exhibit stable behavior within certain neighborhoods. Local linearization around these neighborhoods is a common practice in engineering, particularly because many systems incorporate non-linear elements.

Knowing the output matrix (C) is a common assumption in control design. As the output matrix denotes which linear combination of the internal state is being measured, typically denoted as a sensor. These sensors can degrade and experience health changes as well, but this variable is outside of the specified presenting theorems.

There exist theoretical hard constraints such that the true plant and input matrices must follow a specific decomposition; however, these constraints are required for the proof. Future work could explore methods to relax and bypass some of these constraints, as well as consider the influence of stochastic processes.

The theorems presented do not require feedback from the true system, allowing the estimation technique to be implemented without jeopardizing the true system's stability. Consequently, these estimation schemes can run alongside various input methods used to regulate the system's output. This parallel implementation enables continuous monitoring of changes in the true system's dynamics. If a significant health change occurs, identified through decomposition, technicians and operators can be alerted to inspect and replace affected areas of the system.

Due to the coupled nature of the adaptive laws, adaptive gains depend on one another to change. When combined with asymptotic convergence requirements, the error states can take orders of magnitude of time to converge, as demonstrated in the illustrative example shown in Section V. Tuning the adaptive gains and introducing a fixed gain can reduce the time in which the error state converges to zero by orders of magnitude; however, the amount of time for the error state to converge is still a pressing issue. An attempt had been made to introduce a feedback filter on the adaptive laws as described in *Fuentes's 2025* robust study, as successful implementation can further reduce the convergence time of the error states [2]. However, numerical results were not promising. This motivated the implementation of the feedback filter only on the adaptive input-matrix estimation law. Comparing the illustrative examples in Section V and Section VI reveals a clear difference in the time scales required for the error states and adaptive gains to reach steady state. This further highlights the challenge of incorporating two adaptive mechanisms simultaneously to manage system uncertainty.

Future work will focus on decoupling the adaptive laws by introducing an additional output matrix for the input matrix, which will serve as an encoder for the input and input matrix response. Nevertheless, this approach presents new challenges, particularly concerning the observability conditions for the input matrix. Considering the numerical results, the presented Theorem 1 is likely the best fit for offline use or slow dynamic systems such as structures, whereas Theorem 2 is better suited for faster dynamic systems such as robotics.

VIII. Conclusion

The true system dynamics and input matrix can be influenced by a health status change, resulting in potential changes in constitutive constants and internal interactions. If these changes are not considered in the modeled EOM, discrepancies will emerge between the system model and the true system response. To address the impact of the health status change of the true system's plant and input matrix, a set of coupled adaptive laws were derived. These laws ensure that the error states between the model and the true system converge to zero; however, their effectiveness depends on a specific decomposition. Due to their coupled nature and asymmetric convergence characteristics, the error states may require a significant amount of time to converge. Consequently,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

these laws may not be practical for fast dynamics, such as in robotics, when applied in real-time. Instead, they are more suited for offline applications or for relatively slow dynamic systems, like structures. In contrast, when using only a single adaptive law in the estimator, convergence time can differ by several orders of magnitude, making the approach more suitable for online applications. Future work will focus on relaxing some of the decomposition constraints and exploring methods to decouple the adaptive laws.

References

[1] K. Fuentes, M. Balas, and J. Hubbard, "A control framework for direct adaptive state and input matrix estimation with known inputs for lti dynamic systems," in *Adaptive 2025*, 2025.

[2] K. Fuentes, M. Balas, and J. Hubbard, "A robust control framework for direct adaptive state estimation with known inputs for linear time-invariant dynamic systems," *Applied Sciences*, vol. 15, no. 12, p. 6657, 2025.

[3] D. Luenberger, "An introduction to observers," *IEEE Transactions on Automatic Control*, vol. 16, no. 6, pp. 596–602, 1971.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, Mar. 1960, ISSN: 0021-9223. DOI: 10.1115/1.3662552.

[5] J. C. Doyle, "A review of μ for case studies in robust control," *IFAC Proceedings Volumes*, vol. 20, no. 5, pp. 365–372, 1987.

[6] K. M. Nagpal and P. P. Khargonekar, "Filtering and smoothing in an h/sup infinity/setting," *IEEE Transactions on Automatic Control*, vol. 36, no. 2, pp. 152–166, 1991.

[7] K. Fuentes, M. J. Balas, and J. Hubbard, "A control framework for direct adaptive estimation with known inputs for lti dynamical systems," in *AIAA SCITECH 2025 Forum*, 2025, p. 2796. DOI: 10.2514/6.2025-2796.

[8] T. D. Griffith, M. J. Balas, and J. E. Hubbard Jr, "A modal approach to the space time dynamics of cognitive biomarkers," Ph.D. dissertation, Texas AM University, 2022.

[9] M. Balas and R. Fuentes, "A non-orthogonal projection approach to characterization of almost positive real systems with an application to adaptive control," in *Proceedings of the 2004 American Control Conference*, IEEE, vol. 2, 2004, pp. 1911–1916.

[10] J.-J. E. Slotine and W. Li, *Applied Nonlinear Control*. Prentice hall Englewood Cliffs, NJ, 1991, vol. 199, pp. 123–126.

[11] R. M. Goldwyn, K. Chao, and C. Chang, "Bounded input—bounded output stability of systems," *International Journal of Control*, vol. 12, no. 1, pp. 65–72, 1970.

[12] T. D. Griffith and M. J. Balas, "An adaptive control framework for unknown input estimation," ser. ASME International Mechanical Engineering Congress and Exposition, vol. Volume 7A: Dynamics, Vibration, and Control, Nov. 2021, V07AT07A006. DOI: 10.1115/IMECE2021-67484.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Why the Symbol Grounding Problem Matters in Virtual Reality: A Meme-Focused Solution Based on the Model Human Processor with Real-Time Constraints

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

Makoto Toyota
T-Method
Chiba, Japan
Email: pubmtoyota@mac.com

Katsuko T. Nakahira
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: katsuko@vos.nagaokaut.ac.jp

***Abstract*— Virtual Reality (VR) environment presents objects that the users perceive and interact with. The system then transitions to a new state, reflecting the outcome of the interaction based on the user's perception of the presented objects. For seamless interaction between a Virtual Reality (VR) environment and its users, the meanings assigned to objects by VR application developers must align with the meanings perceived by users. This study proposes a method to enable such seamless interaction by treating the objects presented in VR as symbols and addressing the relationship between the meanings embedded in those symbols and the meanings interpreted by users. This is achieved through the lens of the symbol grounding problem, a well-known challenge in the field of artificial intelligence. Based on the Mode Human Processor with Realtime Constraints (MHP/RT), a cognitive architecture capable of handling action selection in everyday environments, we focus on the idea that users' action selections are influenced by memes—culturally transmitted units of meaning passed down through generations. These memes shape users' understanding of their surrounding world. We suggest that the symbol grounding problem can be addressed by identifying and observing the memes that users draw upon in their daily behavior. We suggest that implementing extracted memes within a Virtual Reality (VR) system can address the symbol grounding problem and enable seamless interaction based on a shared understanding of the symbols mediating between the user and the system.**

Keywords- the symbol grounding problem; meme; virtual reality; artificial intelligence; Model Human Processor with Real-Time Constraints; structured meme theory.

I. Introduction

This article is based on the previous work originally presented in AIVR 2025 [1]. In Section IV, a new Section IV-A is introduced to discuss the process through which "symbols" are generated and structured as memes—a necessary consideration when addressing the symbol grounding problem. Furthermore, Section IV-C explains how securely grounded symbols are utilized in real life contexts and discusses the observability of structured memes.

Humans acquire information about the external world through the five senses and select and execute appropriate actions based on the situation by engaging the Perceptual, Cognitive, and Motor (PCM) processes. In the perceptual process, perceived information is encoded and represented as symbols, enabling thought to occur in the cognitive process. In the thinking process, memory is used to successively transform the symbols into new representations. Part of the outcome of cognition is a representation of an action sequence that can be executed through the motor process. Building on the idea that human intelligence arises from thought processes that manipulate symbols, Newell proposed the Physical Symbol System as theoretical framework for understanding human intelligence [2]. This idea laid the foundation for Soar [3], [4], one of the most well-established cognitive architectures.

The information received by the perceptual process through the sensory organs originates in the real world. Guided by the sequence of actions represented by symbols generated through the PCM processes, actions are executed in the real world, leading to changes in the environment. When attempting to artificially replicate the interactions humans perform in the real world using a physical symbol system, the system must be capable of linking symbolic representations to real-world referents and deriving meaningful understanding from its interactions with the environment.

Achieving this capability is a fundamental challenge in Artificial Intelligence (AI) research, known as the Symbol Grounding Problem (SGP) [5]. The SGP concerns a machine's ability to link its symbolic representations to real-word referents and derive meaningful understanding through its interaction with its environment. Specifically, it addresses how machines can understand and represent the meaning of objects, concepts, and events in the real world. Without the ability to ground symbolic representations in the real world, machines cannot acquire the rich and complex meanings needed for intelligent behavior such as language processing, image recognition, and decision-making. Addressing the SGP is crucial to building machines that can perceive, reason, and act like humans.

Virtual Reality (VR) environments represent one of the contexts in which humans engage in interaction. In such environment, users interact with computer-generated three-dimensional visual scenes or multi-sensory simulations created through modeling and simulation. VR applications immerse the users in a computer-generated environment that closely replicate aspects of the real world. In a VR environment, user interaction proceeds through user-perceivable objects provided by VR applications. The meaning that the user gives to the objects generated by the VR applications determines how the user interacts with the objects. VR applications can achieve seamless interaction by appropriately handling the meanings users assign to virtual objects. This shows that the SGP is central to seamless VR environments. This paper proposes a method to address the SGP in VR.

This paper is organized as follows. Section II introduces a

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

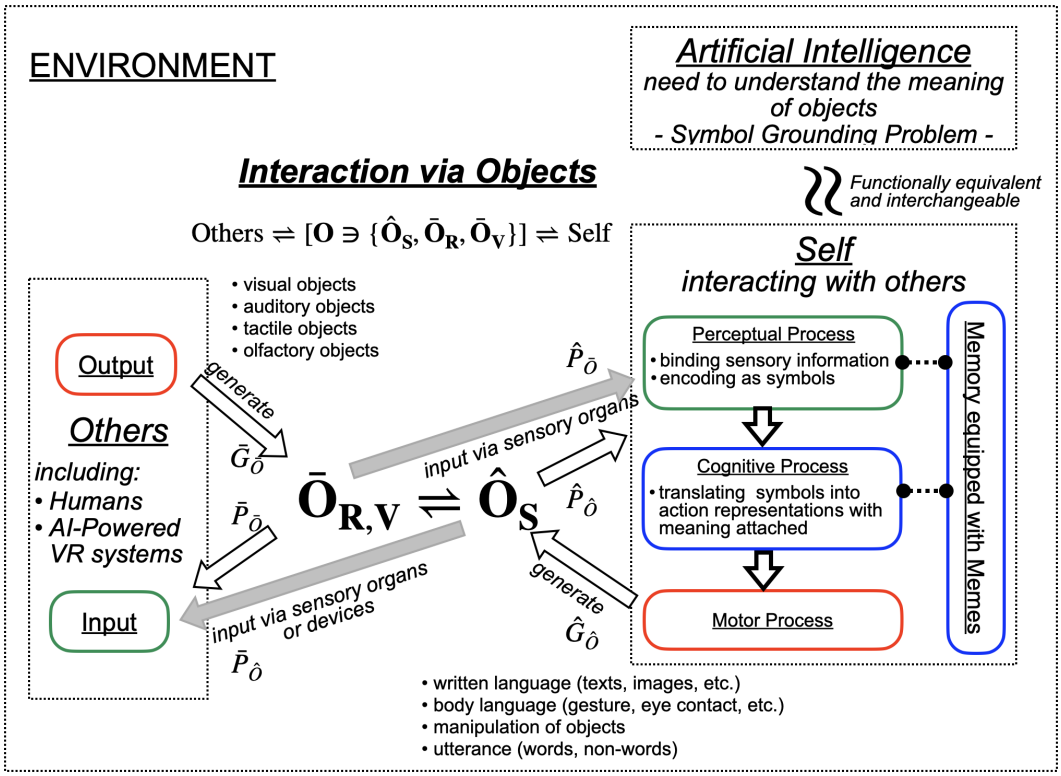


Figure 1. Interaction via objects (adapted from [1, Figure 1])

framework for object-based interaction. Section III examines self–object interaction, approaches to capturing the SGP in AI from an interaction perspective, SGP in AI from the interaction perspective, and how to capture the interaction between self and VR. Section IV proposes methods for generating VR environments that support symbol grounding. Section V concludes this paper by summarizing its findings and outlining future research directions.

II. Interaction via Objects

This section outlines aspects of individual human interaction through objects, with classifications provided in Section II-A and the PCM processes and memory functions described in Section II-B.

A. Interaction Types

Individuals live their daily lives interacting with several objects that exist around them. Objects are classified based on whether the self interacts with them directly or indirectly, and by who created them. There are three types of objects defined as follows:

- $\hat{\mathbf{O}}_{\mathbf{S}}$: Real objects that are directly generated by self, e.g., utterance, written text, drawings, gestures, modeling, etc.
- $\bar{\mathbf{O}}_{\mathbf{R}}$: Real objects generated by other human beings with whom self is directly interacting.
- $\bar{\mathbf{O}}_{\mathbf{V}}$: Objects generated by a machine or other human beings with whom self is not directly interacting.

These three types of objects are collectively represented as $\mathbf{O}$ ($\ni \{\hat{\mathbf{O}}_{\mathbf{S}}, \bar{\mathbf{O}}_{\mathbf{R}}, \bar{\mathbf{O}}_{\mathbf{V}}\}$). In Figure 1, $\mathbf{O}$ is placed in the center, "Self" interacting with $\mathbf{O}$ on its right side, and "Others" interacting with $\mathbf{O}$ on its left side. On top of "Self" is an "Artificial Intelligence" that is functionally equivalent to self and can be replaced. Self operates the PCM processes to generate the objects $\hat{\mathbf{O}}_{\mathbf{S}}$ ($\hat{G}_{\hat{O}}$). Self also perceives them ($\hat{P}_{\hat{O}}$). The objects self perceives include the objects $\bar{\mathbf{O}}_{\mathbf{R,V}}$ generated by others ($\hat{P}_{\bar{O}}$). Meanwhile, others generate the objects $\bar{\mathbf{O}}_{\mathbf{R,V}}$ ($\bar{G}_{\bar{O}}$). If the other is a human being, they run the PCM process, which is equivalent to the one self runs, to generate the objects $\bar{\mathbf{O}}_{\mathbf{R}}$ ($\bar{G}_{\bar{O}_R}$). Machines run their own generative mechanisms to produce the objects $\bar{\mathbf{O}}_{\mathbf{V}}$ as output ($\bar{G}_{\bar{O}_V}$). The input to the others are the objects $\bar{\mathbf{O}}_{\mathbf{R,V}}$ generated by the others themselves ($\bar{P}_{\bar{O}}$), or the objects $\hat{\mathbf{O}}_{\mathbf{S}}$ generated by self ($\bar{P}_{\hat{O}}$). In summary, interaction via objects can be expressed as follows:

$$\text{Others} \rightleftharpoons [\mathbf{O} \ni \{\hat{\mathbf{O}}_{\mathbf{S}}, \bar{\mathbf{O}}_{\mathbf{R}}, \bar{\mathbf{O}}_{\mathbf{V}}\}] \rightleftharpoons \text{Self}$$

B. Interaction between Self and O

Interaction between the self and object $\mathbf{O}$ is driven by the PCM processes, which operate using memory processes that are also updated through PCM execution. This section provides an overview of the PCM and memory processes based on the Model Human Processor with Real-Time Constraints (MHP/RT), a cognitive architecture that can simulate everyday action selections [6]–[8].

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

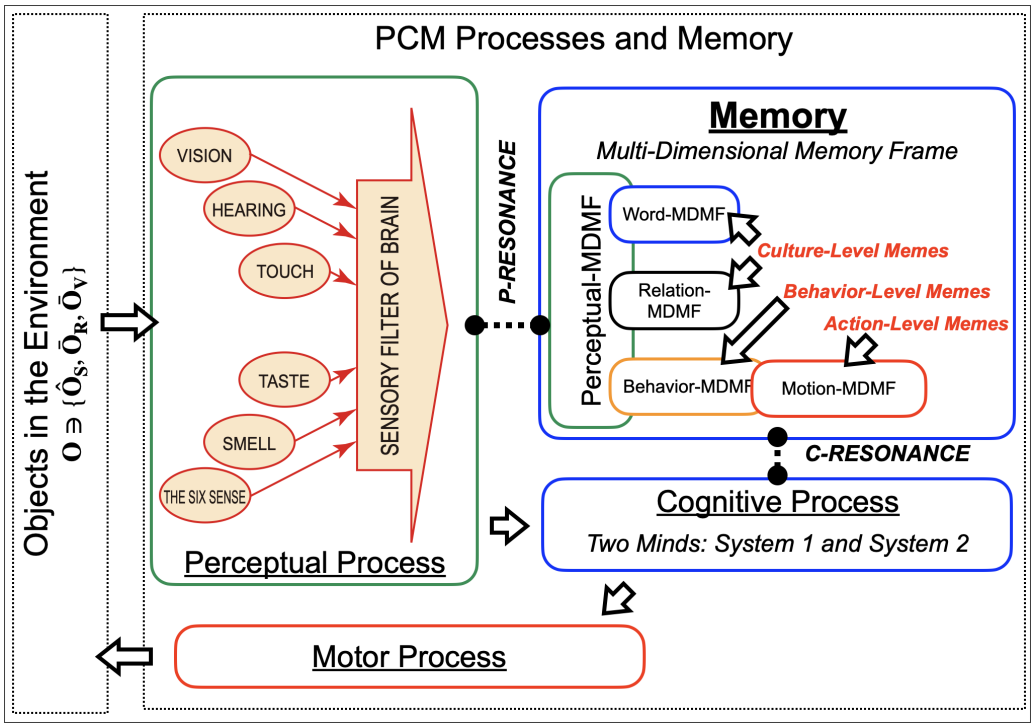


Figure 2. PCM process and memory (adapted from [1, Figure 2])

1) PCM Process: When interacting with objects in the environment, humans receive physical and chemical stimuli through sensory nerves at the body's interface with the environment, allowing environmental information to be processed internally. The brain acquires environmental information concerning the current activity of the self through the multiple sensory organs. Further, it generates bodily movements that are suitable for the current environment. The stable and sustainable relationship between the environment and the self is established through continuous coordination between the activity of the self and the resultant changes in the environment, which should affect the self's next action.

Figure 2, adapted from [9, Figure 1] with some modifications, shows the PCM process, based on the MHP/RT cognitive architecture [7], [8], by which environmental information is taken into the body via sensory nerves, processed in the brain, and then acted upon by the external world via motor nerves. This process uses memory, which is modeled as the Multi-Dimensional Memory Frame. It comprises the Perceptual-, Behavior-, Motor-, Relation-, and Word-Multi-Dimensional Memory Frame. The Perceptual-Multi-Dimensional Memory Frame overlaps with the Behavior-, Relation-, and Word-Multi-Dimensional Memory Frame. This is the unique and indispensable configuration of memories defined by the Multi-Dimensional Memory Frame for spreading activation from the Perceptual- to Motor-Multi-Dimensional Memory Frame, which connects perception with bodily movements.

Perceptual information taken in from the environment through sensory organs *resonates* with information in the Multi-Dimensional Memory Frame, which is called P-Resonance [9]. In Figure 2, this process is indicated by the symbol •—•. Resonance occurs first in the Perceptual-Multi-Dimensional Memory Frame and activates the memory network. After that, the activation spreads to the memory networks that overlap with the Perceptual-Multi-Dimensional Memory Frame, and finally to the Motor-Multi-Dimensional Memory Frame. In cognitive processing based on the Two Minds framework [10], [11], conscious processing (System 2) and unconscious processing (System 1) operate in an interrelated manner [9], [12]. System 2 utilizes the Word- and Relation-Multi-Dimensional Memory Frame via C-Resonance, while System 1 draws on the Behavior- and Motor-Multi-Dimensional Memory Frame via the same mechanism. Motor sequences are then expressed according to the Motor-Multi-Dimensional Memory Frame. The memories involved in the production of actions are updated to reflect the traces of their use process and influence the future action selection process.

2) Memory and Memes: When the PCM process is running, the contents of Perceptual-Multi-Dimensional Memory Frame are updated in response to the perceptual process, those of Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame are updated in response to the cognitive process, and those of Motor-Multi-Dimensional Memory Frame are updated in response to the motor process. Figure 2 illustrates the memories involved in the PCM process—collectively referred to as the Multi-Dimensional Memory Frame—as traces of its operation, and classifies them into five subtypes: Perceptual-, Word-, Relation-, Behavior-, and Motor-Multi-Dimensional

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

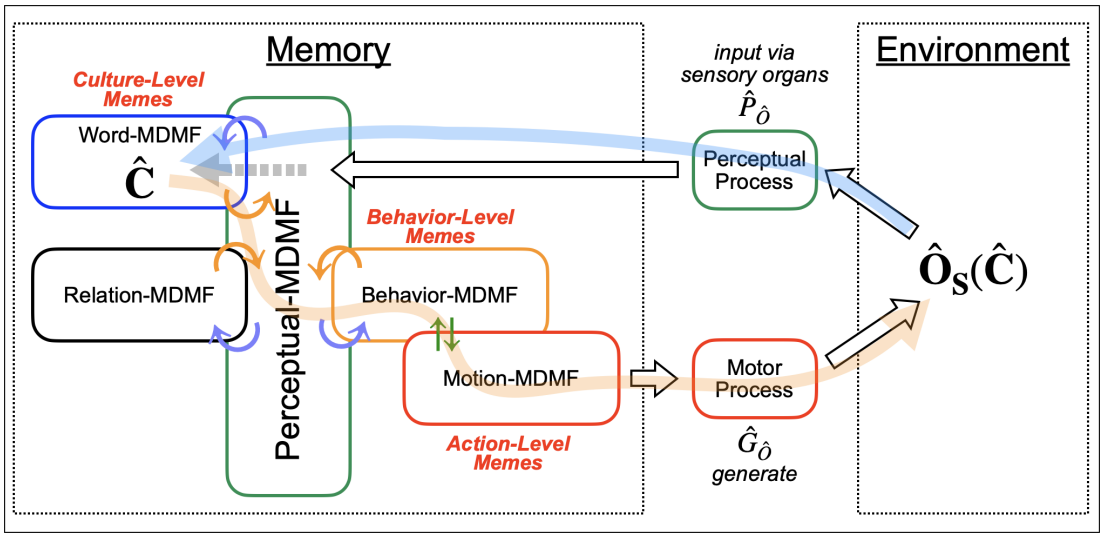


Figure 3. Symbol grounding when the self generates an object $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ that embodies the concept $\hat{\mathbf{C}}$ (adapted from [1, Figure 3])

Memory Frame. Specifically, this is an expression of the way in which the memories are structured, focusing on the continuous updating of memory associated with the execution of PCM process.

Alternatively, the memory system can be viewed from the perspective of memory use. The integrated sensory information first activates the Perceptual-Multi-Dimensional Memory Frame; then the activation spreads to the Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame, and finally to the Motor-Multi-Dimensional Memory Frame bound to the motor nerves. Imitation—doing what one observes—forms the basis of behavior. Therefore, behaviors that can be imitated across generations are preserved as sustainable behaviors. Therefore, we can organize the Multi-Dimensional Memory Frame, which is used by the PCM processes and updated by their execution, regarding *memes* that can be inherited across generations [13].

Word is considered the archetype of meme [14]. Words—also referred to as symbols—are gradually integrated into the environment in stages. First, they appear as thesauruses, which are structured lists of words grouped by synonyms and related concepts. Next, they are incorporated as individual languages used for person-to-person communication, which may involve both literal and metaphorical usage. Finally, they are embedded in cultural languages, where successful communication depends on a shared understanding of the common sense established within a specific community. These three forms circulate among individuals and persist from generation to generation [15].

Thesauruses, individual languages, and cultural languages increase in complexity in that order, based on the patterns through which they are linked to objects in the environment. Thesauruses are associated with the objects in the environment that are encoded in the neural networks in the initial development stage from birth to the age of 3 years. Individual languages are associated with the objects in the environment and the symbols that have already been incorporated in the environment. The same is true for cultural languages. In the Structured Meme Theory (SMT) [13], which is the memory construct to be used and updated by the MHP/RT, the patterns that represent the thesauruses, individual languages, and cultural languages are called action-level meme (A-meme), behavior-level meme (B-meme), culture-level meme (C-meme), respectively.

As shown in the rounded corner rectangle "Memory" in Figure 2, the relationships between the three levels of memes and the Multi-Dimensional Memory Frame are described as follows:

- A-memes represent bodily actions stored in the Motor-Multi-Dimensional Memory Frame.
- B-memes represent behaviors in the environment stored in the Behavior-Multi-Dimensional Memory Frame.
- C-memes represent culture stored in the Relation- and Word-Multi-Dimensional Memory Frame.

III. Details of Interaction between Self and O

This section describes how individual human PCM processes and memory work for each of the interaction types introduced in Section II, and clarifies how the symbol grounding problem is involved and solved.

A. *Interaction between Self and* $\hat{\mathbf{O}}_{\mathbf{S}}$

The process of generating an object $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ that represents the concept $\hat{\mathbf{C}}$ by self ($\hat{G}_{\hat{O}(\hat{C})}$) and perceiving the generated result ($\hat{P}_{\hat{O}(\hat{C})}$) is shown in Figure 3 using Figure 2 as a basis. According to this process, an object that matches the concept is generated. Figure 4 shows the flow of activation within the Multi-Dimensional Memory Frame involved in this process.

This object generation process is indicated along the thick orange arrow in Figure 3. It is denoted symbolically as follows:

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

1) The concept $\hat{\mathbf{C}}$ is represented as a symbol within C-memes, and the nodes associated with it in the Word-Multi-Dimensional Memory Frame are activated.
2) Activation spreads from the Word- to Perceptual-Multi-Dimensional Memory Frame.
3) Activation spreads from the Perceptual- to Relation-Multi-Dimensional Memory Frame that encodes C-memes, and reaches Motor-Multi-Dimensional Memory Frame encoding A-memes via Behavior-Multi-Dimensional Memory Frame encoding B-memes.
4) According to the contents of activated Motor-Multi-Dimensional Memory Frame, the motor process operates to express bodily movements, which result in the generation of object $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ in the environment.

Figure 4. Object generation process (adapted from [1, Figure 4])

Generation Path [G-SS]

$$\hat{\mathbf{C}} \Rightarrow \left[\frac{\text{W-, R-, B-MDMF} \rightleftharpoons \text{M-MDMF}}{\text{P-MDMF}} \right] \Rightarrow \hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}}) \tag{1}$$

Here, the part indicated by $[\cdots]$ in the middle of the expression shows the process of spreading activation in the Multi-Dimensional Memory Frame from the viewpoint of the activated location in the memory. The symbol [G-SS] reads as follows; [**G**enerate - symbol in **S**elf via memory of **S**elf].

The generation process [G-SS] can be rearranged from the perspective of memes, i.e., memories of contents, as follows:

meme mapping

$$\hat{\mathbf{C}} \Rightarrow \left[\frac{\text{C-memes} \rightleftharpoons \text{B-memes}}{\text{P-MDMF}} \rightleftharpoons \text{A-memes} \right]_{\text{M-S}} \Rightarrow \hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}}) \tag{2}$$

The generated object $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ shown on the right end is associated with a state in which the C-, B-, and A-memes activated in the process of spreading activation in the Multi-Dimensional Memory Frame starting from the concept $\hat{\mathbf{C}}$ shown on the left end. The association is represented in the middle, $[\cdots]_{\text{M-S}}$, called "meme mapping" from $\hat{\mathbf{C}}$ to $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$.

1) $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ is perceived and the activation spreads within the Perceptual-Multi-Dimensional Memory Frame.
2) The activation spread from the Perceptual- to Word-Multi-Dimensional Memory Frame results in the activation of symbol related to the perceptual representation.

Figure 5. Object recognition process (adapted from [1, Figure 5])

Figure 5 shows the object recognition process. The thick blue arrow in Figure 3 indicates this process. It is denoted symbolically as follows:

Recognition Path [R-SS]

$$\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}}) \Rightarrow \left[\frac{\text{P-MDFM}}{\text{W-MDMF}} \right] \Rightarrow \hat{\mathbf{C}} \tag{3}$$

The symbol [R-SS] reads [**R**ecognize object - generated by **S**elf using memory of **S**elf]. In [R-SS], if the concept $\hat{\mathbf{C}}$ is strongly activated at the end of the recognition path, then $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ correctly realizes $\hat{\mathbf{C}}$ in the real world. In this case, [G-SS] and [R-SS] are connected and closed, and the symbol $\hat{\mathbf{C}}$ and $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ are cognitively replaceable, which is represented by $\hat{\mathbf{C}} \equiv \hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$. This state can be regarded as a state in which symbol grounding has been achieved within self as shown by Figure 6.

- Symbol $\hat{\mathbf{C}}$ in C-memes activates B- and A-memes via the Perceptual-Multi-Dimensional Memory Frame to generate $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$.
- Perception of $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ activates $\hat{\mathbf{C}}$ and its associated activation pattern in the Multi-Dimensional Memory Frame.
- The perceptual representation of $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ in the Perceptual-Multi-Dimensional Memory Frame is associated with $\hat{\mathbf{C}}$ in Word-Multi-Dimensional Memory Frame.
- In the future, $\hat{\mathbf{C}}$ in the Word-Multi-Dimensional Memory Frame activates perceptual representation of $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ even if it does not exist in the real world, which enables the self to perceptually simulate the concept along with the activation of the Multi-Dimensional Memory Frame necessary to actually generate the object.

Figure 6. Symbol grounding of concept $\hat{\mathbf{C}}$ in self (adapted from [1, Figure 6])

B. Symbol Grounding Problem in Artificial Intelligence

The symbol grounding problem in AI shown in Figure 1 is solved by the fact that an activation pattern equivalent to the activation pattern of memes in the Multi-Dimensional Memory Frame that occurs in the process of generating self's $\hat{\mathbf{O}}_{\mathbf{S}}$ also occurs in AI. By ensuring that the meme mappings occur within AI and self are equivalent, i.e., $[\cdots]_{\text{M-AI}} \equiv [\cdots]_{\text{M-S}}$, AI can be a substitute for self. This is summarized in Figure 7, where "Perceptual-Information-Encoding-in-AI" is the substitute for the Perceptual-Multi-Dimensional Memory Frame of human to perform A/D transformation to input the real world data to the AI system. Since the memes are knowledge passed down from generation to generation, it is considered possible to represent them by symbols. The Perceptual-Information-Encoding-in-AI can also be represented in symbols by encoding environmental information by sensors that perform the same function as sensory organs. The symbol grounding problem in AI is thought to be solved by elucidating memes and implementing them in AI. Methods for elucidating memes are discussed in Section IV-C.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

1) The symbol $\hat{\mathbf{C}}$, which is common to self's, in C-memes activates the Perceptual-Information-Encoding-in-AI as well as the associated C-memes in AI.
2) B-memes in AI are activated via the activated portion of Perceptual-Information-Encoding-in-AI.
3) The part of A-memes in AI that overlap with the activated B-memes in AI is activated.

The steps 1, 2, and 3 constitute the meme-mapping of AI, i.e., $[\cdots]_{M\text{-}AI}$.

4) What is expressed by the activated A-memes in AI is implemented as $\hat{\mathbf{O}}_{\mathbf{AI}}(\hat{\mathbf{C}})$ in the real world via appropriate actuators.
5) Upon input of the object $\hat{\mathbf{O}}_{\mathbf{AI}}(\hat{\mathbf{C}})$ in AI, activation spreads in the Perceptual-Information-Encoding-in-AI, followed by the activation of the symbol $\hat{\mathbf{C}}$ in C-memes in AI.
6) The part in the Perceptual-Information-Encoding-in-AI that corresponds to $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ and the symbol $\hat{\mathbf{C}}$ integrate the C-, B-, and A-memes in AI activated in the steps 1, 2, and 3 to form an integrated association. At this point, $\hat{\mathbf{C}} \equiv \hat{\mathbf{O}}_{\mathbf{AI}}(\hat{\mathbf{C}})(= \hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}}))$ is established, in other words, the symbol $\hat{\mathbf{C}}$ both the AI and the self commonly recognize has been grounded, guaranteed by the relationship $[\cdots]_{\text{M-AI}} \equiv [\cdots]_{\text{M-S}}$.

Figure 7. Symbol grounding of concept $\hat{\mathbf{C}}$ in AI (modified from [1, Figure 7])

C. Interaction between Self and $\bar{\mathbf{O}}_{\mathbf{R}}$

Consider the case where the other human generates an object. The object generation process for the concept $\bar{\mathbf{C}}$ performed by the other human is represented as $\bar{\mathbf{C}} \Rightarrow [\cdots] \Rightarrow \bar{\mathbf{O}}_{\mathbf{R}}(\bar{\mathbf{C}})$, which is a the-other-human's version of [G-SS]. The other human spreads activation in his/her own Multi-Dimensional Memory Frame. The meme mapping used for the other human is denoted as $[\cdots]_{\text{M-O}}$. The self interacting with the other human who has just generated $\bar{\mathbf{O}}_{\mathbf{R}}(\bar{\mathbf{C}})$, which could be perceivable objects such as a gesture, utterance, or drawing, perceives and recognizes it according to the following path:

Recognition Path for the Object Generated by Other [R-OS]

$$\bar{\mathbf{O}}_{\mathbf{R}}(\bar{\mathbf{C}}) \Rightarrow \left[\frac{\text{P-MDFM}}{\text{W-MDMF}}\right] \Rightarrow \hat{\mathbf{C}} \tag{4}$$

If the symbol $\bar{\mathbf{C}}$ $(\equiv \bar{\mathbf{O}}_{\mathbf{R}}(\bar{\mathbf{C}}))$ held by the other human and the symbol $\hat{\mathbf{C}}$ held by the self are identical, the symbol is successfully transmitted through the object expressed by the other. For example, the other person holds a certain word $\bar{\mathbf{C}}$ in his/her mind and expresses it physically through gestures, and then the self sees it and assigns the word $\hat{\mathbf{C}}$ to it. The latent word of the other is connected to the self's latent word through the physical actions of the other person. Consider the case of communication via words, where the self and the other look at a sequence of words $\tilde{\mathbf{C}}$. The self and the other perform symbol grounding according to their respective generation paths; $[\cdots]_{\text{M-S}}$ and $[\cdots]_{\text{M-O}}$ are included in each symbol grounding process. If the self and the other have grown up in the same environment, which is the necessary condition for them to have a common set of memes, then the relation $[\cdots]_{\text{M-S}} \equiv [\cdots]_{\text{M-O}}$ would hold, and the shared symbols have the same meaning. However, in the case of $[\cdots]_{\text{M-S}} \neq [\cdots]_{\text{M-O}}$, the meaning of all visually shared symbols may not be shared. For example, the phrase "see you on the ground floor" may trigger different behaviors depending on the culture to which the reader of the phrase belongs.

D. Interaction between Self and $\bar{\mathbf{O}}_{\mathbf{V}}$

In interactions between the self and other humans, the interaction is symmetric because both parties are human. Specifically, in the section represented by $[\cdots]$, which connects symbols and objects along the generation and recognition pathways, activation spreads within the Multi-Dimensional Memory Frame possessed by both the self and the other individual. Conversely, in the case of Self–VR interaction, the generation and recognition paths on the system side are different from those on the human side. In the generation path, symbols defined in the system are converted into objects that can be perceived by the user. In the recognition path, human-generated objects are input to the system via sensors and converted into symbols that the system can handle. Both conversions are performed by programs implemented in the system.

In a VR environment, the system takes in the information emitted by human users and then determines the response to it. In any case, the input is represented as a symbol $\tilde{\mathbf{C}}$. Within the system, after setting a symbol $\bar{\bar{\mathbf{C}}}$ to be transmitted in the next cycle of interaction, the symbol–object transformation is performed and the object $\bar{\bar{\mathbf{O}}}_{\text{V}}(\bar{\bar{\mathbf{C}}})$ is output to be perceived by the user. This generation path is denoted as [G-VV].

The user perceives the object $\bar{\bar{\mathbf{O}}}_{\text{V}}(\bar{\bar{\mathbf{C}}})$ and recognizes it as a symbol along the recognition path [R-VS]. Let $\hat{\mathbf{C}}'$ be the recognized symbol. The user activates his/her Multi-Dimensional Memory Frame along the generation path [G-SS] for $\hat{\mathbf{C}}'$ and obtains the corresponding object $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}}')$. If the relation $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}}') \equiv$ (or $\approx$) $\bar{\bar{\mathbf{O}}}_{\text{V}}(\bar{\bar{\mathbf{C}}})$ holds, the interaction will proceed smoothly. If not, it will fail.

IV. Method of $\bar{\mathbf{O}}_{\mathbf{V}}$ Generation with Symbol Grounding Secured

Only if the object generation path [G-VV] in the system is executed according to the user's meme mapping $[\cdots]_{\text{M-S}}$, that is, if the meme mapping in the system is based on $[\cdots]_{\text{M-AI}}$, it is possible to proceed with an interaction that guarantees symbol grounding. Since memes are knowledge that are passed down from generation to generation, they can be represented by symbols. This section proposes a method to implement meme mapping in the system. To this end, we first explain the process by which new symbols are assigned to patterns activated within the Multi-Dimensional Memory

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Frame during interaction with the external world. These patterns are then repeatedly used in communication with others—a process through which they become memes. Finally, based on this process, we add explanations for the A-, B-, and C-memes introduced in Section II-B2, and suggest a method for externalizing them by observing users' daily behavior.

A. Memeification of Patterns

Figure 8 depicts how memes propagate in the reality field [15]. The process of propagation is facilitated by "symbolization." Using Figure 8, we will look at the process of meme propagation, focusing on the process of symbolization.

1) Resonance with Objects in the Environment: Suppose that Object-O_1, Object-O_2, or Object-O_3 appear in the environment (bottom-left in Figure 8). These objects are elements of $\mathbf{O} \ni \{\hat{\mathbf{O}}_\mathbf{S}, \bar{\mathbf{O}}_\mathbf{R}, \bar{\mathbf{O}}_\mathbf{V}\}$ shown in Figure 1. Here, it does not matter how the objects were generated. It is simply a situation in which they appear in front of self. Each resonates with the patterns of elements in the Perceptual-Multi-Dimensional Memory Frame of the self to activate them. This process is depicted as P-Resonance in Figure 2. Subsequently, activation propagates through the Multi-Dimensional Memory Frame through chain firings in the memory to activate the Relation-, Behavior-, and Motor-Multi-Dimensional Memory Frame. The activated portion of memory, which is shown as Pattern-P_a at the bottom of Figure 8, will be used to select next actions in System 1 and System 2 cognitive processes through C-Resonance as shown in Figure 2. Pattern-P_a is associated with different types of patterns, namely, patterns of elements of objects in the Perceptual-Multi-Dimensional Memory Frame, patterns of relations in the Relation-Multi-Dimensional Memory Frame, or patterns of temporal changes in the Behavior- and Motor-Multi-Dimensional Memory Frame.

2) Mapping Environment on Brain: The entities in the environment that resonate with Pattern-P_a, i.e., Object-$O \ni \{\text{Object-}O_1, \text{Object-}O_2, \text{Object-}O_3, \cdots\}$, are taken as a single object, Object-$O[P_a]$, in the action selection process conducted by System 1 and System 2 as shown at the bottom-right of Figure 8.

3) Mapping Patterns on Word-Multi-Dimensional Memory Frame (Symbolization): Once a unique name, Symbol-S_a, is attached to Pattern-P_a, Symbol-S_a is stored in the Word-Multi-Dimensional Memory Frame while maintaining the relationship with Pattern-P_a. In Figure 8 this is shown by a thick double-headed arrow connecting bottom-right and top-left areas as mapping of pattern on the Word-Multi-Dimensional Memory Frame. Symbol-S_a might be associated with any of the elements in Object-O, i.e., $\{\text{Object-}O_1, \text{Object-}O_2, \text{Object-}O_3, \cdots\}$. It is the process by which a large amount of information in a pattern is aggregated into a single symbol. By aggregating more information, the range of what can be expressed by a symbol is expanded [16].

4) Importing Symbol-S_a *from the Word-Multi-Dimensional Memory Frame to Environment:* Once Pattern-P_a is symbolized as Symbol-S_a in the self's Word-Multi-Dimensional Memory Frame, it becomes a part of the environment, which is shown as an arrow from top-left to bottom-left of Figure 8. Namely, the relationship, Symbol-$S_a \in \mathbf{O}$, is established. This makes possible the following situation to emerge; when Symbol-S_a appears as an object in the environment, it resonates with the Perceptual-Multi-Dimensional Memory Frame via P-Resonance to activate Pattern-P_a, followed by activation of patterns of relation in the Relation-Multi-Dimensional Memory Frame and temporal changes in the Behavior- and Motor-Multi-Dimensional Memory Frame to cause C-Resonance. In Section III-A, on the assumption that the the concept, $\hat{\mathbf{C}}$, exists, the generation path of the object representing it, $\hat{\mathbf{O}}_\mathbf{S}(\hat{\mathbf{C}})$, is illustrated (see Figure 3 and equation (1)). Alternatively, in this section, when an object $\mathbf{O}(*)$ of unknown generator with no symbol attached within self appears in the environment, a new symbol Symbol-S_a is assigned to $\mathbf{O}(*)$. As a consequence, when the self recognizes the symbol in the environment which is available to anyone who is in it, the object, $\hat{\mathbf{O}}_\mathbf{S}(\text{Symbol-}S_a) \equiv \mathbf{O}(*)$, can be generated along the generation path shown by equation (1). Since Symbol-S_a is created by starting from the generator unknown object, $\mathbf{O}(*)$, object $\mathbf{O}(*)$ could be recognized as Symbol-S_a. The symbol Symbol-S_a is grounded on the real object $\hat{\mathbf{O}}_\mathbf{S}(\text{Symbol-}S_a) \equiv \mathbf{O}(*)$.

5) Utilization of the Word-Multi-Dimensional Memory Frame in Language Behavior: The processing flow explained so far is centered on Pattern-P_a and therefore is closed in self and the environment. This situation corresponds to the one described in Section III-A. However, Symbol-S_a incorporated in self's Word-Multi-Dimensional Memory Frame is available for interaction with others described in Section III-C, which is communication or language behavior depicted in top-right of Figure 8. Language behavior emerges via successive activation of the Word-, Relation-, Behavior-, and Motor-Multi-Dimensional Memory Frame, and utilization of them by System 1 and System 2 via C-Resonance. The resonated part of memory corresponds to Object-$O[P_a]$, which is the mapping of Pattern-P_a, generated via P-Resonance with the environment, to the brain.

In the situation where Symbol-S_a incorporated in the self's Word-Multi-Dimensional Memory Frame, which is denoted as Symbol-S_a(self) to indicate explicitly it belongs to the self, is utilized in communication with others, Symbol-S_a(self) must connect to Symbol-S_a(other) incorporated in the other's Word-Multi-Dimensional Memory Frame. These symbols are apparently identical and refer to a unique symbol AbstractSymbol-AS_a. In Figure 8, the relationship between AbstractSymbol-AS_a used in communication and Object-$O[P_a]$ directly connected with the self's symbol Symbol-S_a(self) is indicated by a thick double-headed arrow from bottom-right to top-right.

6) Maintenance and Inheritance of Abstract Concepts: The single abstract concept, AbstractSymbol-AS_a, is referred to Symbol-S_a(self) in the self's Word-Multi-Dimensional Memory Frame and Symbol-S_a(other) in the other's Word-Multi-Dimensional Memory Frame, which are denoted as $\hat{\mathbf{C}}$ and $\bar{\mathbf{C}}$, respectively, in Section III-C. Only when the meme mappings, $[\cdots]_{\text{M-S}}$ and $[\cdots]_{\text{M-O}}$, intervening in the respective symbol

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

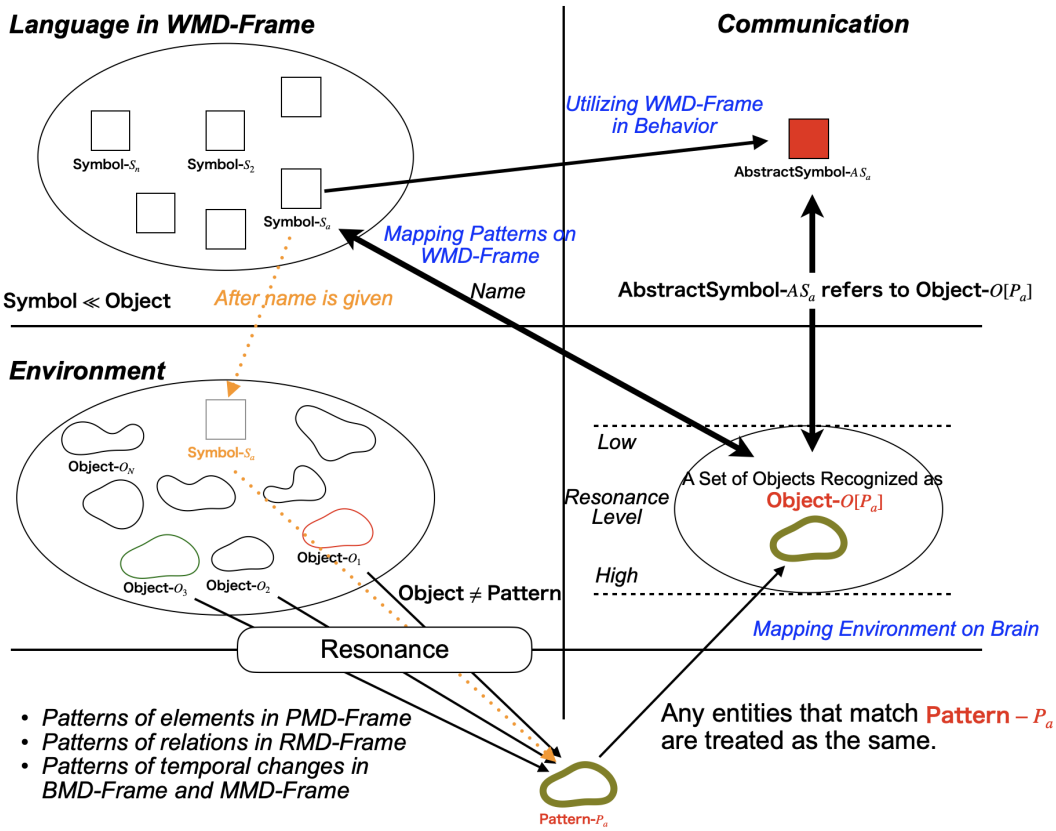


Figure 8. Memeification of patterns ([15, Figure 4] modified)

groundings of self and other are congruent, communication through AbstractSymbol-$AS_a \equiv \{\hat{\mathbf{C}}, \bar{\mathbf{C}}\}$ can be established while maintaining a common symbol grounding for both parties. When communication proceeds in this way, abstract concept AbstractSymbol-AS_* is maintained and inherited. This is the reality of the meme. It is reasonable to assume that the higher the resonance level of Symbol-S_a becomes, the longer AbstractSymbol-AS_a is maintained and inherited as an effective communication medium.

B. Getting Memes into the Brain

As explained in Section IV-A, a new symbol is generated within the Word-Multi-Dimensional Memory Frame. It becomes grounded as a real-world element through its association with a set of real-world objects and circulates within the behavioral ecology—including self and others—through interpersonal communication. It is a necessary condition for circulation that the symbols used in the community have the same symbol grounding. Those that circulate within the behavioral ecology are called memes. As each individual grows, the content of memes that are acquired changes. As explained in Section II-B2, first, the action-level memes involved in the basic action are acquired. Next, memes at the behavior level, which are sequences of actions, are acquired. Finally, the cultural-level meme—forming the background that shapes and supports actions and behaviors—is acquired.

Figure 9 shows an enlarged version of Pattern-P_a displayed at the bottom center of Figure 8. The processes shown in Figure 8 as "Mapping Environment on Brain" and "Mapping Patterns on Word-Multi-Dimensional Memory Frame" are shown in the lower right and upper right of Figure 9, respectively. The process of "Mapping Patterns on Word-Multi-Dimensional Memory Frame" can be subdivided into the three types described above according to the developmental stage, as shown in Figure 9, depending on the degree of complexity of the mapping [13].

1) Action-Level Memes: During the period from birth to two to three years of age, humans generate several synapses that connect neural circuits in the brain and take in as much information as possible. The rate of synapse generation is then reduced, and the distribution of information up to that point is used to determine the basic characteristics of the sensory organs. Simultaneously, by initiating body movements and imitating the movements of the individuals around them, they acquire body movements that have been formed empirically and accumulatively as individual ecology. This is formed through life's skillful method of adjusting the growth of muscles and other parts of the body to external constraints. Simultaneously,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

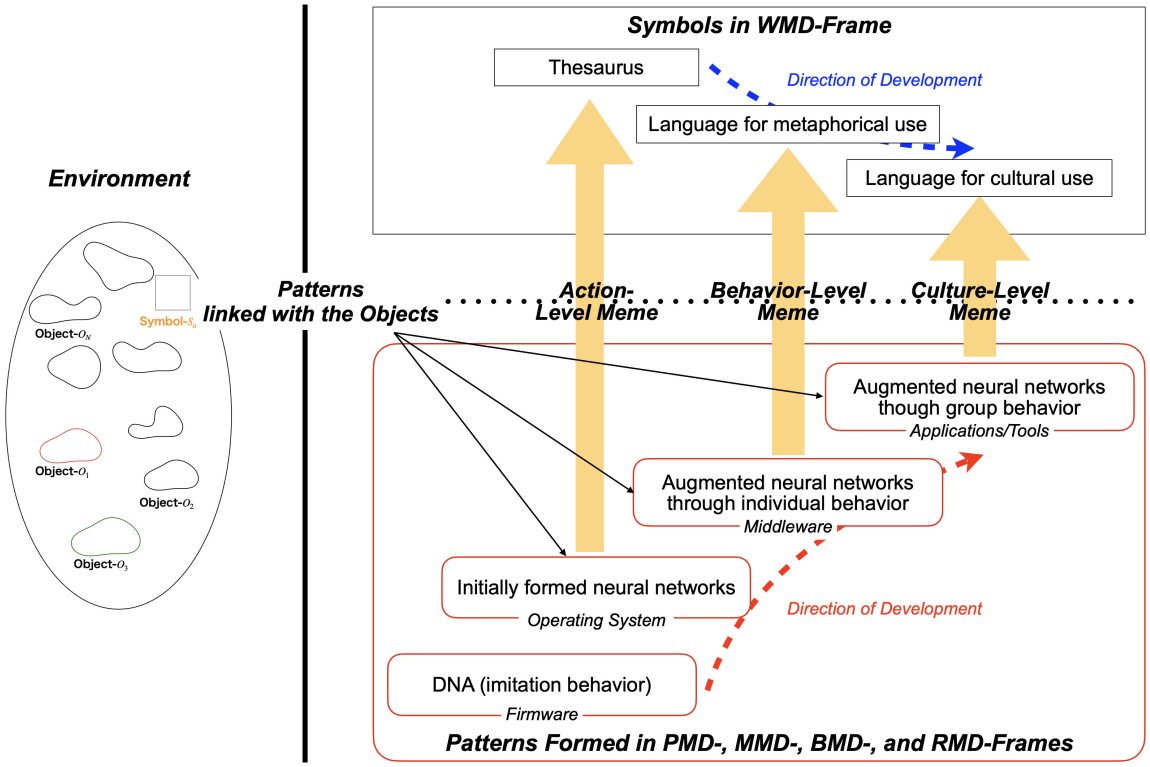


Figure 9. Structure of meme (adapted from [15, Figure 5])

information from the sensory organs is linked to bodily actions. The most important bodily functions formed at this stage are the voice and hand functions.

2) Behavior-Level Memes: Later, the voice paves the way to speech; the hands pave the way to tool use. Through continuous imitation, humans learn to use the words and tools of those around them. At this time, humans acquire a new hierarchy of actions by organizing and summarizing the fact that a particular collection of sounds evokes a particular response, and that the feel of a hand experienced through tactile sensation and the movement of a tool perceived visually are captured as an unified whole via the tool. This is achieved by linking the A-memes formed on the brain circuit according to the situation in which they are used, and making them available as a coherent whole.

3) Culture-Level Memes: Furthermore, words pave the way to language, and tools pave the way to the use of more complex machines. At this stage, humans learn to act as members of the culture and civilization of the group to which they belong, through imitation and through the experience of autonomous activities as members of the group. At this time, the B-memes is extended to be used in a complex manner, and culture-specific behavior patterns are formed.

C. Utilizing Memes in Behavioral Ecology with Realtime Constraints

Figure 3 shows the situation in which symbol grounding is realized. This includes the process of starting with the concept, initiating the activated action, and the generation of object. However, this process can occur without exchanging the Perceptual-Multi-Dimensional Memory Frame and Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame shown in Figure 3—that is, without ensuring existence through a link to perceptual reality. In practice, generating objects while ensuring their existence is inefficient and may therefore be omitted in real-life situations [12]. Objects that are guaranteed to be symbol grounded are often used in communication without explicitly verifying their grounding—particularly in interactions constrained by real-time demands. Under this situation, the meme mapping shown in equation (2) can be rewritten as follows:

$$\hat{\mathbf{C}} \Rightarrow \left[\frac{\text{C-memes} \rightleftharpoons \text{B-memes}}{\textit{no reference to P-MDMF}} \rightleftharpoons \text{A-memes} \right]_{\text{M-S}} \Rightarrow \hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}}) \quad (5)$$

D. Extraction of Memes through Mapping Memes into Information Systems

Since memes are symbols, they can be expressed by language. Any elements appearing in the equation (5) are observable in the language. It is possible to extract A-, B-, and C-memes by organizing the verbal descriptions of actions performed top-down in daily life at the three levels of memes without referring to the Perceptual-Multi-Dimensional Memory Frame.

The mechanism by which the memes inherit information is

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

analogous to an information system. This is shown in bottom-right of Figure 9. A-memes serve as the operating system that defines general patterns of spatial-temporal behavioral functions. B-memes represent middleware that extends the general patterns to concrete patterns. C-memes act as application tools that extend the concrete patterns to the ones that work in a number of groups of individuals.

The internal processing of information systems is represented by symbols. A necessary condition for the system and the user to achieve common symbol grounding is that $[\cdots]_{\text{M-AI}} \equiv [\cdots]_{\text{M-S}}$ holds. When trying to describe the memes contained in the contents of $[\cdots]_{\text{M-S}}$ in equation (5), the fact that an analogy can be made between memes at each level and the system mechanism indicates that the approach of describing memes by mapping them to information systems is effective. Representation of memes in the behavioral ecology in which the VR system is implemented regarding information system's symbols could be transferred to representation of $[\cdots]_{\text{M-AI}}$ that is equivalent to $[\cdots]_{\text{M-S}}$, opening the way for system implementation with guaranteed symbol grounding.

E. Extraction of Memes: Examples

We have attempted to understand people's behavioral selection in various situations by assuming that it is performed according to MHP/RT, in the form of "under such-and-such a circumstance, such-and-such a person with such-and-such an attribute will express such-and-such a behavior." How a person acts reflects the functioning of the PCM process, and what the person does—the specific content of their actions—relates to the memes stored in their Multi-Dimensional Memory Frame and those activated during the execution of the actions. Cognitive Chrono-Ethnography (CCE) has been developed as a methodology for understanding behavior using this approach and has been applied to various situations [17].

Meme extraction has been attempted in some studies focusing on particular behavioral domains;

1) Navigation behavior that allows elderly people to reach their destinations without getting lost while obtaining necessary information from information displays in unfamiliar train stations [6];
2) The proficient creative activity that is the emanation of memes as the traditional skill of the ceramic artist, including memes that form the traditional skill that is handed down [18];
3) Proficient piano players practicing in preparation for competition performance [19].

It is expected that memes involved in interaction within VR environments can also be elucidated using the two-wheeled approach defined by the theory—MHP/RT—and its practical methodology, CCE, by consulting previous studies. By representing the memes activated within the VR environment using the vocabulary of the information system, it becomes possible to simulate the symbol grounding performed by the user. This enables the construction of a VR environment that connects seamlessly with the user in a forward-looking manner.

V. Conclusion

Whether in the real world or in VR, individual humans execute the actions they select by running their PCM processes in response to objects that appear in the interacting world. This study posits that smooth, seamless interaction requires shared object meanings among interacting agents. We explored the concept of "symbol grounding," as discussed in artificial intelligence, within the context of user interaction in VR. AI will not achieve symbol grounding unless the same structured memes as humans are put into AI. In this study, while demonstrating that memes are observable, a path to how this is possible was presented.

We argued that symbol grounding is realized when we perceive and recognize an object generated from a certain symbol and match it with that symbol. In the object generation process, A-, B-, and C-memes, are involved in the conversion from symbols to objects. Since memes are passed down from generation to generation, we suggested that they could be extracted by observation, referring to the previous studies [6], [18], [19]. These are based on the MHP/RT [6]–[8] and the SMT [13], [15]. The ability to evaluate objects that are presented when users interact with VR environments from the symbol grounding perspective is an important issue in the development of seamless VR environments.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 19K12246 / 19K12232 / 20H04290 / 22K12284 / 23K11334, and National University Management Reform Promotion Project. The authors would like to thank Editage (www.editage.com) for the English language editing.

References

[1] M. Kitajima, M. Toyota, and K. T. Nakahira, "Addressing the Symbol Grounding Problem in VR", in *AIVR 2025 : The Second International Conference on Artificial Intelligence and Immersive Virtual Reality*, 2025, pp. 56–62.

[2] A. Newell, "Physical Symbol Systems", *Cognitive Science*, vol. 4, no. 2, pp. 135–183, 1980. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog0402_2.

[3] A. Newell, *Unified Theories of Cognition (The William James Lectures, 1987)*. Cambridge, MA: Harvard University Press, 1990.

[4] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An architecture for general intelligence", *Artificial Intelligence*, vol. 33, pp. 1–64, 1987.

[5] S. Harnad, "The symbol grounding problem", *Physica D: Nonlinear Phenomena*, vol. 42, no. 1, pp. 335–346, 1990, [retrieved: 2, 2025], ISSN: 0167-2789. DOI: https://doi.org/10.1016/0167-2789(90)90087-6.

[6] M. Kitajima and M. Toyota, "Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT)", *Behaviour & Information Technology*, vol. 31, no. 1, pp. 41–58, 2012. DOI: 10.1080/0144929X.2011.602427.

[7] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)", *Biologically Inspired Cognitive Architectures*, vol. 5, pp. 82–93, 2013.

[8] M. Kitajima, *Memory and Action Selection in Human-Machine Interaction*. Wiley-ISTE, 2016, ISBN: 9781848219274.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[9] M. Kitajima *et al.*, "Basic Senses and Their Implications for Immersive Virtual Reality Design", in *AIVR 2024 : The First International Conference on Artificial Intelligence and Immersive Virtual Reality*, 2024, pp. 31–38.

[10] D. Kahneman, "A Perspective on Judgment and Choice", *American Psychologist*, vol. 58, no. 9, pp. 697–720, 2003.

[11] D. Kahneman, *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux, 2011.

[12] M. Kitajima, M. Toyota, J. Dinet, and K. T. Nakahira, "Transforming Conscious Goals into Unconscious Actions in Real-world Interactions: Real-world Use of Behavioral Ecological Memes via GOMS", *International Journal On Advances in Intelligent Systems*, vol. 18, no. 3 & 4, in press, 2025.

[13] M. Toyota, M. Kitajima, and H. Shimada, "Structured Meme Theory: How is informational inheritance maintained?", in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008, p. 2288.

[14] D. C. Dennett, *From Bacteria to Bach and Back: The Evolution of Minds*. W W Norton & Co Inc, Feb. 2018.

[15] M. Kitajima, M. Toyota, and J. Dinet, "How Resonance Works for Development and Propagation of Memes", *International Journal On Advances in Systems and Measurements*, vol. 14, pp. 148–161, 2021.

[16] M. Kitajima *et al.*, "Language and Image in Behavioral Ecology", in *COGNITIVE 2022 : The Fourteenth International Conference on Advanced Cognitive Technologies and Applications*, 2022, pp. 1–10.

[17] M. Kitajima, "Cognitive Chrono-Ethnography (CCE): A Behavioral Study Methodology Underpinned by the Cognitive Architecture, MHP/RT", in *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, Cognitive Science Society, 2019, pp. 55–56.

[18] M. Kitajima, M. Toyota, and J. Dinet, "Art and Brain with Kazuo Takiguchi", in *COGNITIVE 2023 : The Fifteenth International Conference on Advanced Cognitive Technologies and Applications*, 2023, pp. 1–10.

[19] K. T. Nakahira, M. Kitajima, and M. Toyota, "Understanding Practice Stages for a Proficient Piano Player to Complete a Piece: Focusing on the Interplay Between Conscious and Unconscious Processes", *International Journal on Advances in Life Sciences*, vol. 16, pp. 164–177, 2024.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Transforming Conscious Goals into Unconscious Actions in Real-world Interactions: Real-world Use of Behavioral Ecological Memes via GOMS

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

Makoto Toyota
T-Method
Chiba, Japan
Email: pubmtoyota@mac.com

Jérôme Dinet
Université de Lorraine
Nancy, France
Email: jerome.dinet@univ-lorraine.fr

Katsuko T. Nakahira
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: katsuko@vos.nagaokaut.ac.jp

***Abstract*— Our daily actions are executed to achieve desired states. Perceiving our own situation, we select actions that are expected to realize the desired states and successively execute them. The memory used in this process is a representation in the brain in the form of memes that are passed on from generation to generation. Memes are structured into three levels—action, behavioral, and cultural—and are acquired through mimetic behavior. Higher-level memes are acquired as one gets older. This study draws on the Model Human Processor with Real-time Constraints (MHP/RT), a cognitive architecture that includes perceptual, cognitive, and motor (PCM) processes, and a memory system that is used during action selection by the PCM process and updated after action execution. We examine how the cognitive process of Two Minds utilizes memes structured in the three layers, termed C-resonance in MHP/RT. Meanwhile, knowledge built as a result of iterative actions toward a goal state is represented by a GOMS hierarchical structure comprising goals (G), operators (O), methods (M), and selection rules (S). We show that GOMS bundles memes belonging to different levels, combining goals and selection rules at the conscious level with methods and operators at the unconscious level to achieve effective and efficient goal-oriented action execution. The expressed behavior can be regarded as the result of crossing the syntax expressed by GOMS with the semantics expressed by memes, showing distinct characteristics depending on the balance of dominance between unconscious and conscious behaviors in the behavioral ecology. By examining the behavioral ecology from the perspective of GOMS, we can see how static memes are implemented in a dynamic behavioral ecology and how behaviors with guaranteed corporality are expressed.**

Keywords- GOMS; Behavioral ecology; Meme; Resonance; Corporeality.

I. Introduction

This work is based on a study previously presented at COGNITIVE2025 [1]. The basics of the concept of the process of developing perceptual, cognitive, and motor processes, and forming memories while interacting with the real world were added to Section I-A to reinforce the introduction of Figures 3 and 4, which provide the basis of this work. In addition, we added Figure 6 to Section III-B3, referring to the role of GOMS in the process of embodying "words", a typical example of memes. Furthermore, a discussion of the relationship between the nature of the happiness goals at the top of the goal structure and the GOMS that develops the behavioral goals underneath it was added to Section IV-B3 as Section IV-B3a.

A. Development and Memory Formation through Real-World Interaction

What we observe as each individual's physical behavior is the result of multiple processing with a Parallel Distributed Processing (PDP) system [2], [3] and not with a single unified system. This PDP system is organized evolutionally and realized as a neural network system, including the brain, the spinal nerves, and the peripheral nerves.

According to Damasio [4], a vertebrate animal develops its neural network system in the following way. It starts with the development of the paired structure comprising the sense of touch and its associated reflex movements. Then, the senses of smell and taste, and finally, sight and hearing develop their associations with reflex movements. From the beginning, the perceptual stimuli from the five senses form a paired structure with their associated reflexive movements. In addition, the association tends to become bidirectional to establish "selective sensing," which is a paired structure with feedback between perception and movement. For example, the sense of hearing and motor movements for vocalization establish a feedback loop between them immediately after one acquires the function of voicing.

In summary, the neural network system first forms the autonomic nervous system of the respective autonomous organs shown in the center of Figure 1 as a genetic fundamental structure. Then, it is crossed with the somatic nervous system that controls reflexive movements associated with the perceptual stimuli from the five senses. Finally, it develops feedback loops with a system of interneurons that connect these systems. Figure 1 schematically depicts this loop.

Each individual lives in the environment by operating the perceptual, cognitive, and motor (PCM) processes along this loop. They then systematically develop a cross-network of sensory, motor, and intervening nervous systems [7]. The genealogy of the DNA of vertebrates suggests that Perception, Interneuron, and Motion—i.e., PIM—form the basis for de-

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

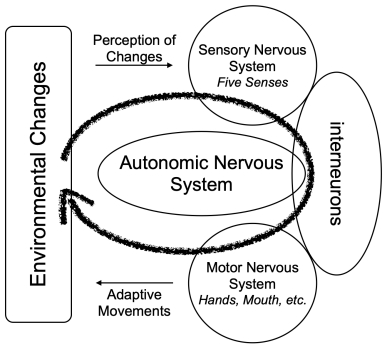


Figure 1. Continuous cyclic loop of perception and movement (adapted from [5, Figure 1])

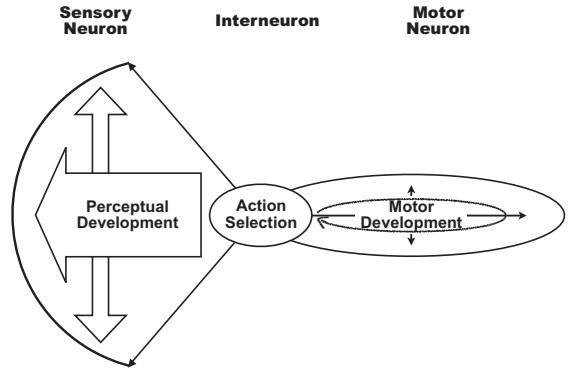


Figure 2. Development of the sensory nervous system and the somatic nervous system, and interneurons connecting them with action selection process (adapted from [6, Figure 1])

velopment. Figure 2 shows that PIM is the basis of the formation of the body based on neural circuits and the whole body is formed as PIM develops. Perception (P) captures various kinds of environmental changes via the sensors with different properties (M-dimension). Motor movement (M) is performed continuously and cyclically from birth, and develops gradually through increasing precision and optimizing intensity (N-dimension). The interneurons (I) memorize the effective interlocking relationships between P and M ($M \otimes N$ mapping), and construct neural circuits that function in a feedforward fashion and which allow for more complex feedback control to enhance the effectiveness of reactions. PIM develops by expanding the behavioral-ecological bandwidth. The behavioral ecological categories of vertebrates, such as acquiring food and raising children, are almost identical and within a limited range. Everyday human life is performed cyclically in a behavioral ecological band which is expanded by performing new actions to realize a goal within the limited behavioral ecological categories and adding these to the existing band.

While interacting with the environment for living, humans develop by selecting and executing actions, and accumulating execution results while operating The PCM processes along the loop are shown in Figure 2. *The basis of action selection is imitation.* The things—any cultural entity including objects and events that an observer might consider a replicator of a certain idea or set of ideas—to be imitated exist as memes; they are repeatedly imitated from generation to generation.

B. Meme Proposed by Dawkins [8]

The mechanism by which the cultures and civilizations produced by humans are passed on from generation to generation was not clear. From the standpoint of cultural anthropology, Dawkins organized his research on the mechanisms of cultural inheritance. He argued that cultural inheritance cannot be explained solely in terms of the capability of memory on the part of humans, and that there must be a hypothetical existence on the part of culture that might convey information, such as genes. He coined the term "meme" for this indefinite virtual entity [8]. This idea itself received substantial support, but time passed without the mechanism being clarified [9].

When Dawkins proposed the meme, the function of genes was not yet understood. Therefore, Dawkins' explanation and that of others had many problems inherent in them due to misunderstandings about genes. Certainly, genes were replicators. However, they did not play the role of duplicating the blueprint of the finished product as conventionally thought. Rather, they played the role of plotting the process of growth that established the basic functional structure and its relationships. It was this role of genes that enabled humans to be highly adaptable.

C. Redefining Memes through their Connection to the PCM Process

The memes proposed by Dawkins can be redefined by considering them as mappings of the individual's memory (which can be called *the individual ecological memes*), which is activated in the process of selecting and executing actions, onto *the collective ecology* that caries the culture. Memes are realized in each individual's memory. They hold the relationships between events, which enable humans in an ever-changing environment to express effective behavior in each situation in generic forms that are valid across generations [10]. Specifically, the spatial coordinates and absolute times that characterize events occurring in the real world are *not* retained in the memes; they are dynamically determined according to the state of the environment when the behavior is expressed according to the representation of the memes.

Figure 3 shows a schematic representation of the action selection and execution process in the real environment as a PCM process. In the perceptual process, humans perceive the state of the real environment through parallel processing of the five senses and integrate what they individually perceive by binding them. In the cognitive process, memes related to the perceived information are activated in parallel and integrated as a series of operators that can be executed as concrete actions in the environment. In the motor process, the operators are executed through feedforward processing, keeping

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

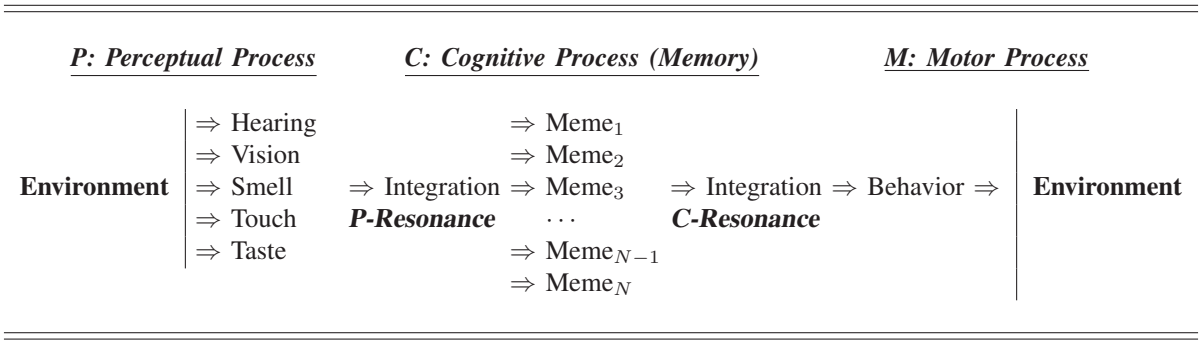


Figure 3. PCM process and memes

pace and synchronizing with changes in the environment in an unconscious manner.

D. Problem Statement

The PCM process synchronously runs with the environment. Meanwhile, the "memory system" used by the PCM process updates asynchronously itself with the environment to reflect the results of the PCM process. Supported by the PCM process and memory system, each individual repeats action selections in an ever-changing environment without any breakdown. Here, we must clarify the interface between the PCM process, which operates synchronously with the environment, and memory system, which is not required to synchronize with the environment but is connected to the PCM process. The mechanism for connecting what is perceived with memory has been described as P-resonance [11]. Within the memories, structured with memes as elements, activation propagates in parallel with the integrated perceptual information as the activation source. Connecting activated memories to the actions performed in the real world can be rephrased as "enabling the activated memes in the real world by integration." In Figure 3, this mechanism is shown as C-resonance. How is this done?

When we unravel the origin and evolution of life, we can find a clue to the solution. Life is formed under the structures shaped by the atmosphere, oceans, energy cycles, and gravity that characterize the Earth, a planet in our solar system, spinning on its own axis and orbiting the sun. The direction of life's evolution is determined by the pressures exerted by these structures. Life is formed as an adaptive body with the functional and structural features that work most efficiently in the environment. This is best captured by the four elements of Goals (G), Operators (O), Methods (M), and Selection (S) rules (GOMS) [12]. Here, we show that C-resonance, which integrates the memes activated in parallel as the effective actions in the real world, can be explained by the GOMS concept.

E. Organization of the Article

This remainder of this article is organized as follows. Section II describes the PCM process and memes, referring to our previous work. Section III describes the GOMS theory presented by Card, Moran, and Newell in their book, "The Psychology of Human-Computer Interaction," and describe the mechanism of C-resonance. Section IV discusses the characteristics of the behavioral ecology that emerges from the interaction of the structured meme content and C-resonance by GOMS. Section V summarizes this study and discusses its implications for the digital generation.

II. PCM Process and Memes

Here, we explain the details of the PCM process, memories, and memes outlined in Figure 3, based on our cognitive architecture "Model Human Processor with Real-time Constraints (MHP/RT)" [13], [14].

A. PCM Process and Resonance for Linking with Memory

When interacting with the environment, humans respond to physical and chemical stimuli emitted from the external and internal environment via the sensory nerves located at the interface with the environment, and take in environmental information in the body. The brain acquires environmental information concerning the self's current activity through the multiple sensory organs. Further, it generates bodily movements suitable for the current environment. The stable and sustainable relationship between the environment and self is established through continuous coordination between the activity of the self and resultant changes in the environment, which should affect the self's subsequent action.

Figure 4, adapted from [11, Figure 1], shows the process, based on MHP/RT [13], [14] via which environmental information is taken into the body via sensory nerves, processed in the brain, and then acted upon by the external world via motor nerves. This process involves memory, which is modeled as Multi-Dimensional Memory Frame, and perceptual, cognitive (Two Minds), and motor processes. The memory structure, Multi-Dimensional Memory Frame, comprises Perceptual-, Behavior-, Motor-, Relation-, and Word-Multi-Dimensional Memory Frame. Perceptual-Multi-Dimensional Memory Frame overlaps with Behavior-, Relation-, and Word-Multi-Dimensional Memory Frame for spreading activation from Perceptual- to Motor-Multi-Dimensional Memory Frame.

The perceptual information taken in from the environment through sensory organs resonates with information in the memory network structured as Multi-Dimensional Memory Frame; that is, P-Resonance. In Figure 4, this process is

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

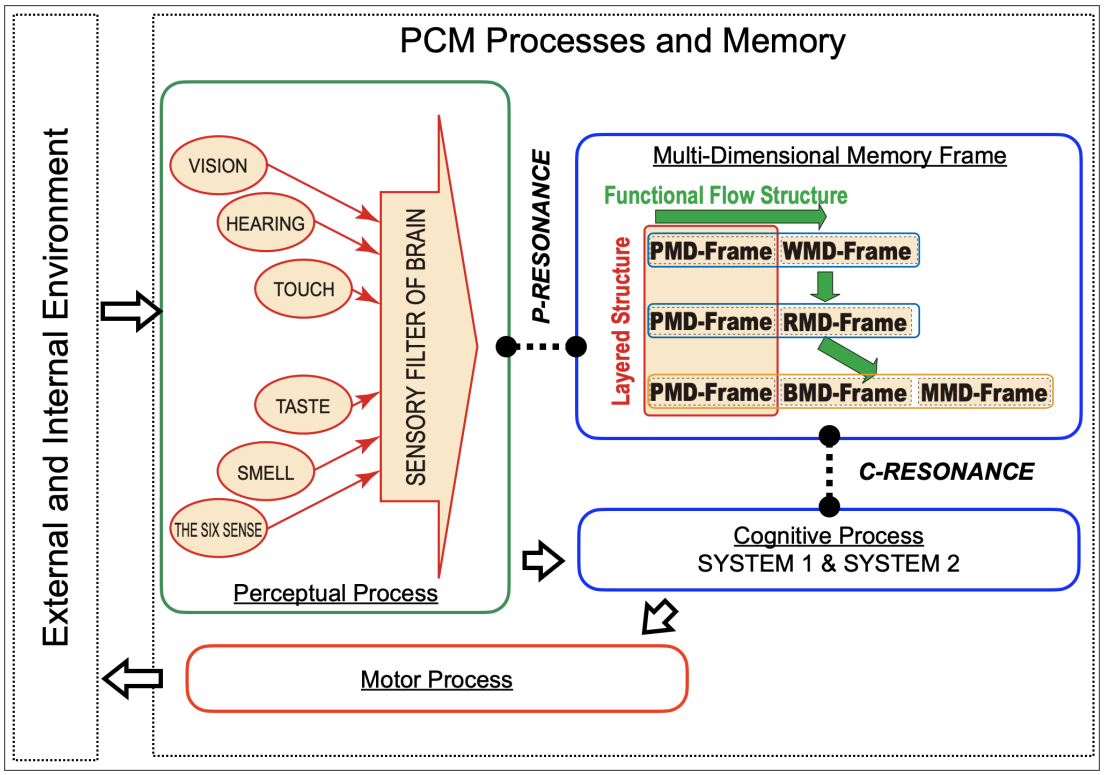


Figure 4. Information uptake by perceptual processes from the external and internal environments' memory activation, and execution of cognitive and motor processes through resonance (adapted from [11, Figure 1])

indicated by •—•. Resonance occurs first in the Perceptual-Multi-Dimensional Memory Frame and activates the memory network. Then, the activity propagates to the memory networks that overlap the Perceptual-Multi-Dimensional Memory Frame, which are Behavior-, Relation-, and Word-Multi-Dimensional Memory Frame, and finally to Motor-Multi-Dimensional Memory Frame. In cognitive processing by Two Minds, conscious processing by System 2—which utilizes the Word- and Relation-Multi-Dimensional Memory Frame via C-Resonance—and unconscious processing by System 1—utilizing the Behavior- and Motor-Multi-Dimensional Memory Frame via C-Resonance—proceed in an interrelated manner. The motor sequences are expressed according to the Motor-Multi-Dimensional Memory Frame, which is the result of cognitive processing. The memories involved in the production of actions are updated to reflect the traces of its use process and influence the future action selection process.

This feature of MHP/RT should be contrasted with the goal-oriented cognitive architectures such as ACT-R [15], [16] in which the conscious processes are considered as the processes to control people's behavior and the unconscious processes are considered subordinate to the conscious or intentional processes. What ACT-R tries to do is to show how System 2 can be implemented on top of System 1. The procedural memory system is very similar to System 1 (fast, learning based on rewards/experience, intuitive), and then ACT-R models tend to consist of a set of production rules that—when run on this System 1 module and in combination with symbolic working memory buffers and a long-term memory system—give rise to the slower, deliberative planning behaviors seen in System 2. This is a very different approach to that given in this paper. However, ACT-R models are totally adequate for simulating stable human activities with weak time constraint in which deliberate decision making would work effectively, but might be hard for the situations with strong time constraint where the environmental condition changes chaotically and deliberate decision making implemented on System 2 might not work as effective [13].

B. P-Resonance Connecting Perceptual Processes and Multi-Dimensional Memory Frame

Information from the environment is taken into the brain via multiple sensory organs. The sensory organs are distributed throughout the body. In addition, they receive time-series information. That is, this is spatially and temporally distributed. The brain integrates this disparate sensory information in some way, perceives it, and passes it on to cognitive processing. The question of how this integration is performed is known as the binding problem. We proposed that P-resonance provides a solution to the binding problem and showed the existence of basic senses that enable the orderly processing of information from sensory organs. The basic senses include the rhythmic sense related to time, spatial sense related to spatial perception, and number sense [11].

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

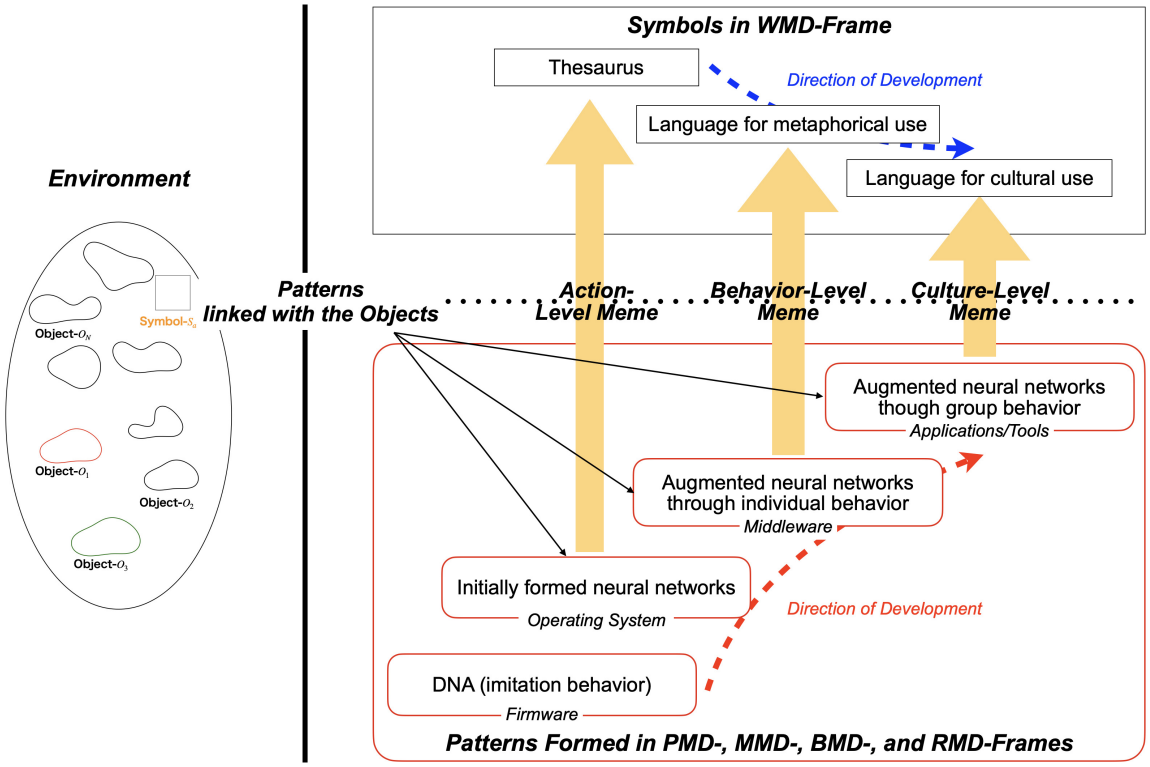


Figure 5. Structure of the meme (adapted from [10, Figure 5])

C. *Memory as a Structured Meme*

When the PCM process is running, the contents of Perceptual-Multi-Dimensional Memory Frame are updated in response to the perceptual process, those of Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame are updated in response to the cognitive process, and those of Motor-Multi-Dimensional Memory Frame are updated in response to the motor process. Figure 4 characterizes the memories of PCM process—i.e., the Multi-Dimensional Memory Frame—as the traces of its operation and classifies it into five sub-memories: Perceptual-, Word-, Relation-, Behavior-, and Motor-Multi-Dimensional Memory Frame. In short, it expresses how the memories are structured, focusing on the continuous updating of memory associated with the execution of PCM process. Notably, in the Multi-Dimensional Memory Frame, Perceptual-Multi-Dimensional Memory Frame overlaps with Behavior-, Relation-, and Word-Multi-Dimensional Memory Frame, for spreading activation from Perceptual-Multi-Dimensional Memory Frame to Motor-Multi-Dimensional Memory Frame.

Alternatively, the memory system can be viewed from the perspective of memory use. The integrated sensory information through the basic senses first activates the Perceptual-Multi-Dimensional Memory Frame (P-resonance); then, the activation propagates to the Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame, and finally, to the Motor-Multi-Dimensional Memory Frame bound to the motor nerves. This process is repeated in an environment that changes from moment to moment. Satisfactory behavior is expressed in the environment. The basis of behavior is imitation; however, what is imitable is limited according to growth stage. In addition, behaviors that can be imitated across generations are preserved as sustainable behaviors. In this way, we can organize the Multi-Dimensional Memory Frame, which is used by the PCM processes and updated by their execution, in terms of memes that can be inherited across generations [17].

Figure 5 shows a functional classification of the regions of the Multi-Dimensional Memory Frame that are activated by P-resonance after an object in the environment has been perceived. “Words” are considered the archetype of meme [18]. Words (i.e., symbols) in the Word-Multi-Dimensional Memory Frame are gradually incorporated into the environment in the form of *thesauruses* (i.e., lists of words in groups of synonyms and related concepts); languages used for person-to-person communication, *individual languages*, which might include not only direct but also metaphorical uses; and languages used in cultural contexts, *cultural languages*, in which appropriate understanding of common sense that has been established in the specific community. Words are essential for successful communication.

Thesauruses, individual languages, and cultural languages increase their complexity in this order in terms of the patterns linked with the objects in the environment. Thesauruses are associated with the objects in the environment which are encoded in the neural networks in the initial development

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

stage from the birth to three years. Individual languages are associated with not only the objects in the environment but also the symbols already incorporated in the environment. The same is true for cultural languages.

The process of "Mapping patterns on symbols in Word-Multi-Dimensional Memory Frame" can be subdivided into three processes based on the degree of complexity of mapping. The patterns that were mapped on the thesauruses, individual languages, and cultural languages are shown as Action-, Behavior-, and Culture-Level Memes, respectively, which were introduced in the Structured Meme Theory proposed by Toyota et al. [17]. Hereafter, the memes classified into these memes are abbreviated by A-, B-, and C-memes, respectively.

The culture that exists in the environment is the integration of the C-memes which exist in the brain of an individual across all members of the group to which that individual belongs. This corresponds to the meme proposed by Dawkins. Blackmore, one of the theoretical followers of Dawkins' meme theory, argues that "memes are symbolized by the act of imitation" after examination of the theory of the meme [19]. This argument is consistent with the idea of memes presented here, since we can think of it as saying that the A- and B-memes are physical and provide a stable basis for imitation, and that the C-memes above them is not mentioned because it is strongly dependent on the environment.

The mechanism by which the three levels of memes, and genes inherit information is analogous to an information system. Genes serve as firmware that mimics behavior-level activities. A-memes serve as the operating system that defines general patterns of spatial-temporal behavioral functions. B-memes represent middle-ware that extends the general patterns to concrete patterns. C-memes act as application tools that extend the concrete patterns to the ones which work in a number of groups of people.

The relationships between the three levels of memes and Multi-Dimensional Memory Frame are as follows:

- A-memes represent bodily actions stored in the Motor-Multi-Dimensional Memory Frame,
- B-memes represent behaviors in the environment stored in the Behavior-Multi-Dimensional Memory Frame, and
- C-memes represent culture stored in the Relation- and Word-Multi-Dimensional Memory Frame.

Meme-based behaviors (i.e., mimicry behaviors) are implemented in the real environment. Since what can be imitated depends on the individual growth stage, there are qualitatively different sets of mimicry behaviors to emerge depending on the growth stage. The bases of those mimicry behaviors are represented as A-, B-, and C-memes.

As Dawkins proposed, taking a meme-centric view of cultural inheritance is in itself essential. A person's genes express a memory resonance response mechanism, through which replications (resonance replications) are generated when there is a common experience. Memes are present in the environment as those which can cause these resonance replication. Such memes can be called cultures. Memes influence the phenotype called culture. However, the resonance itself is formed as something unique in a person's experience. The fact that imitation is personal and influenced by environmental conditions does not guarantee that the phenotype will be perfectly imitated. This mechanism is common to the idea of ecological inheritance theory, named niche (ecological status) construction, advanced by Odling-Smee et al. [20]. This mechanism allows humans to be highly adapted to their environment.

III. C-Resonance via GOMS

A. Binding Problem at the Cognitive Level

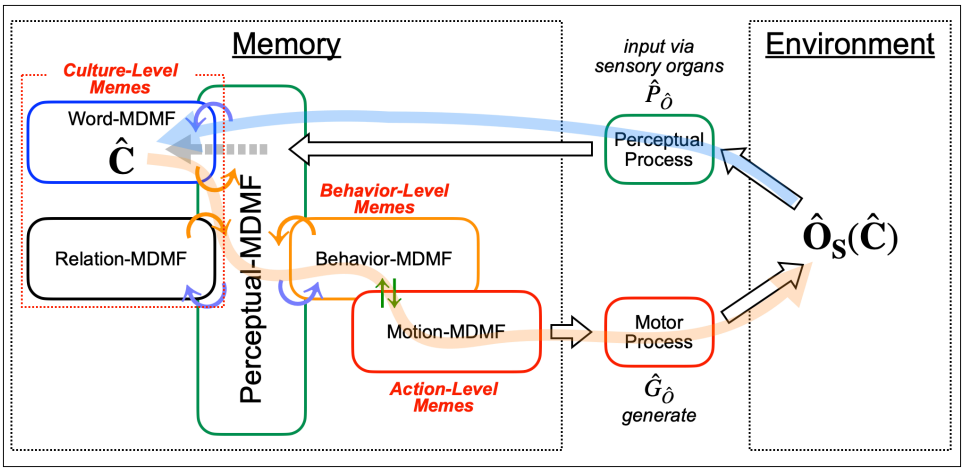
In Figure 5, the objects in the environment activate the A-, B-, and C-memes. In this activation process, various regions related to the objects are activated. In Figure 4, the propagation of activation within the Multi-Dimensional Memory Frame is shown as the functional flow structure. The expression is such that the activity propagates from the Perceptual-Multi-Dimensional Memory Frame to Word-, Relation-, Behavior-, and Motor-Multi-Dimensional Memory Frame in that order. However, the layers below the Word-Multi-Dimensional Memory Frame are not structurally overlapped. Therefore, the activity propagates layer by layer from the top to bottom via the Perceptual-Multi-Dimensional Memory Frame that overlaps with them; at the top, there is an activation flow from the Word- to Perceptual-Multi-Dimensional Memory Frame; at the middle, from the Perceptual- to Relation-Multi-Dimensional Memory Frame; and at the bottom, from the Perceptual- to Behavior-, and finally, to Motor-Multi-Dimensional Memory Frame. The portions of Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame which are activated in this manner may contain multiple regions that may be related via the Perceptual-Multi-Dimensional Memory Frame but not directly related to each other.

Memories that hold memes are activated in parallel to be used by the PCM process, which is a serial process. Here, we can see another binding problem occurring at the cognitive level. In Figure 4, the bridge between cognitive and memory processes is shown as C-resonance for resolving the meme binding problem. The cognitive process may operate *carefully* by using the entire areas of the Multi-Dimensional Memory Frame that are activated in connection with the Perceptual-Multi-Dimensional Memory Frame. The advantage of this method is that reality can be guaranteed by referring to the contents active in the Perceptual-Multi-Dimensional Memory Frame. However, it is *inefficient* because it is an interpreter-like process. At the perceptual level, the binding problem of perceptual information is solved by P-resonance for the effective use of the perceptual information. At the cognitive level, C-resonance resolves the problem of efficient use of memory by binding memes somehow that are activated in parallel [21]. What is the equivalent of the basic senses in P-resonance in C-resonance?

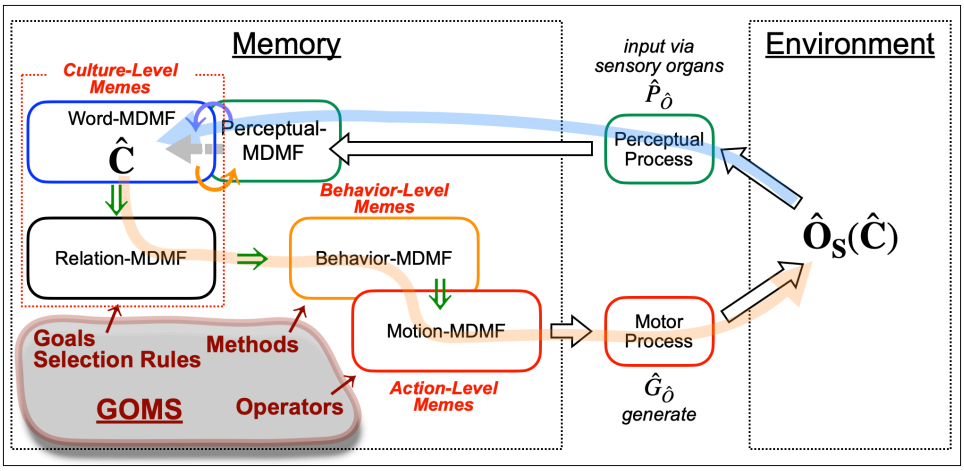
B. GOMS

1) GOMS as a Meta-structure for Understanding Behavioral Ecology: Humans select actions help realize the state they desire to achieve. The principles at work while executing

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a)



(b)

Figure 6. (a) Symbol grounding when the self generates an object $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ that embodies the concept $\hat{\mathbf{C}}$ (adapted from [22, Figure 3]). Goal is a kind of $\hat{\mathbf{C}}$, and $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ represents the state in the environment showing that "the goal is accomplished."; (b) Symbol grounding for object generation—i.e., goal accomplishment—using GOMS which *does not* involve Perceptual-Multi-Dimensional Memory Frame

Activation of Memes via P-Resonance			**Utilization of Memes via C-Resonance**	
Perceptual Process	Multi-Dimensional Memory Frame	Memes	GOMS	Cognitive Process
	Word- & Relation-	Culture-Level	Goals and Selection Rules	System 2
Basic Senses	Behavior-	Behavior-Level	Methods	System 1
	Motor-	Action-Level	Operators	System 1

Figure 7. Relation between GOMS, memes, and Two Minds

the cognitive activity of action selection are bounded rationality and satisficing principle [23]. Such action selection processes are modeled by the serial firing of procedural knowledge expressed in the form of production rules: "IF conditions are satisfied, THEN perform actions." As a cognitive architecture for modeling such situations, ACT-R exists [15], [16], [24]. Individual action choices are expressed in terms of firing sequences of production rules, which are procedural knowledge. However, if we take a bird's-eye view of the situations that each individual encounters, the firing sequences of procedural knowledge applied in more or less similar situations will show certain patterns. Card, Moran and Newell [12] identified GOMS as a concept that represent such patterns. GOMS specifies concepts that define a meta-structure which is essential for understanding the ecology of human behavior. Aristotle's theory of the four causes was the first theory to systematize such a meta-structure. Allen Newell et al. reconstructed it from a cognitive scientific perspective and constructed the GOMS theory.

2) Definition of GOMS: GOMS is an analytic technique for making quantitative and qualitative predictions about

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

skilled behavior with a computer system. GOMS is defined as follows (adapted from [12, Chapter 5, pp.144–146]):

> The user's cognitive structure consists of four components: (1) a set of Goals, (2) a set of Operators, (3) a set of Methods for achieving the goals, and (4) a set of Selection rules for choosing among competing methods for goals. We call a model specified by these components a GOMS model.
>
> ***Goals***. A goal is a symbolic structure that defines a state of affairs to be achieved and determines a set of possible methods by which it may be accomplished.
>
> ***Operators***. Operators are elementary perceptual, motor, or cognitive acts, whose execution is necessary to change any aspect of the user's mental state to affect the task environment.
>
> ***Methods***. A method describes a procedure for accomplishing a goal. It is one of the ways in which a user stores this knowledge of a task. The description of a method is cast in a GOMS model as a conditional sequence of goals and operators, with conditional tests on the contents of the user's immediate memory and on the state of the task environment.
>
> ***Selection Rules***. When a goal is attempted, the user may have more than one available method to accomplish the goal. The selection of which method to be used need not be an extended decision process. It may be that the task environment features dictate that only one method is appropriate. Meanwhile, a genuine decision may be required. The essence of skilled behavior is that these selections are not problematical, and that they proceed smoothly and quickly without the eruption of puzzlement and search that characterizes problem-solving behavior. In a GOMS model, method selection is handled by a set of selection rules. Each selection rule is of the form "if such-and-such is true in the current task situation, then use method M."

Behavioral goals are represented by a robust hierarchical structure. There is a primary behavioral goal, G. Underneath it, there are subgoals, G′, that must be accomplished to complete the primary goal. Finally, there are sub-subgoals, G″, to complete the individual subgoals. The final node that undertakes the task is the operator. One node above it is the method, while one level above it is the node representing the selection rule. The goal structure from top to bottom looks like "G-G′-G″ $\cdots$ S-M-O."

3) Binding Memes via GOMS: In GOMS, behavioral goals form a robust hierarchical structure. The goal structure mediates the organization of behavior. Achieving a goal, G, requires achieving the subgoals underneath it, G′s. This structure does not hold the time as its primary parameter. The order between G′s is important. The time elapsed for executing G′ is associated with the operators located at the bottom layer, which connect to the motor process of PCM that implements the contents of Motor-Multi-Dimensional Memory Frame—i.e., the operators—in the real world.

Meanwhile, the mechanism of action execution based on MHP/RT is explained as follows. As shown in Figure 4, the environment is perceived and connected to the Multi-Dimensional Memory Frame by P-resonance. Then, as shown in Figure 3, the memes acquired by structuring the Multi-Dimensional Memory Frame through experience are activated, while A-memes are connected to the real world to execute the action. As mentioned earlier, in the functional flow structure within the Multi-Dimensional Memory Frame shown in Figure 4, behavior generation following the flow of activity through the Perceptual-Multi-Dimensional Memory Frame is inefficient. This process is shown in Figure 6(a). For the concept $\hat{\mathbf{C}} \in$C-meme that the self wants to realize in the real world, the activation propagates from the Word-Multi-Dimensional Memory Frame to Relation- and Behavior-Multi-Dimensional Memory Frame via Perceptual-Multi-Dimensional Memory Frame, and finally reaches the Motor-Multi-Dimensional Memory Frame. The goal expressed in language here is a kind of concept. Then, the object $\hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ is realized in the real world through the motor process. Here, the state in which the goal has been achieved is realized; it exists as a perceptual object in the environment. This object is perceived by the self and Perceptual-Multi-Dimensional Memory Frame is activated. The activation propagates to the Word-Multi-Dimensional Memory Frame and the object is recognized as $\hat{\mathbf{C}}$, $\hat{\mathbf{C}} \equiv \hat{\mathbf{O}}_{\mathbf{S}}(\hat{\mathbf{C}})$ is established and the envisioned concept is realized in the real world. In other words, the envisioned goal has been realized in the environment. The establishment of this relationship also means that the symbol grounding problem [25], which has been a challenge in the field of artificial intelligence, has been solved [22].

One may reasonably assume that GOMS is used to structure A-, B-, and C-memes which do not contain absolute temporal and spatial information as a method of realizing behavior generation without breaking down, while keeping in sync with the real world where the situation changes from moment to moment. GOMS should correspond to the phenomenon of A-, B-, and C-memes binding without the Perceptual-Multi-Dimensional Memory Frame when encountering certain situations, indicating the entity of the phenomenon of C-resonance. This may correspond to a shortcut that may be formed within the Multi-Dimensional Memory Frame.

Figure 7 shows the correspondence between memes and GOMS. Among the activated memes, the combinations of C-, B-, and A-memes which have formed GOMS bonds in the process of gaining experience are processed by System 2 and System 1, while the operator sequence is executed in the real world [26]. Figure 6(b) shows how GOMS provides a shortcut by substituting the mapping from the Word- to Motor-Multi-Dimensional Memory Frame, where the symbol grounding shown in Figure 6(a) is guaranteed.

IV. Discussion: Deepening the Understanding of Behavioral Ecology

A. GOMS-construct Structure and Behavioral Ecology

1) GOMS-construct Structure: Figure 8 shows the overall GOMS-construct, which has been constructed by experience,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

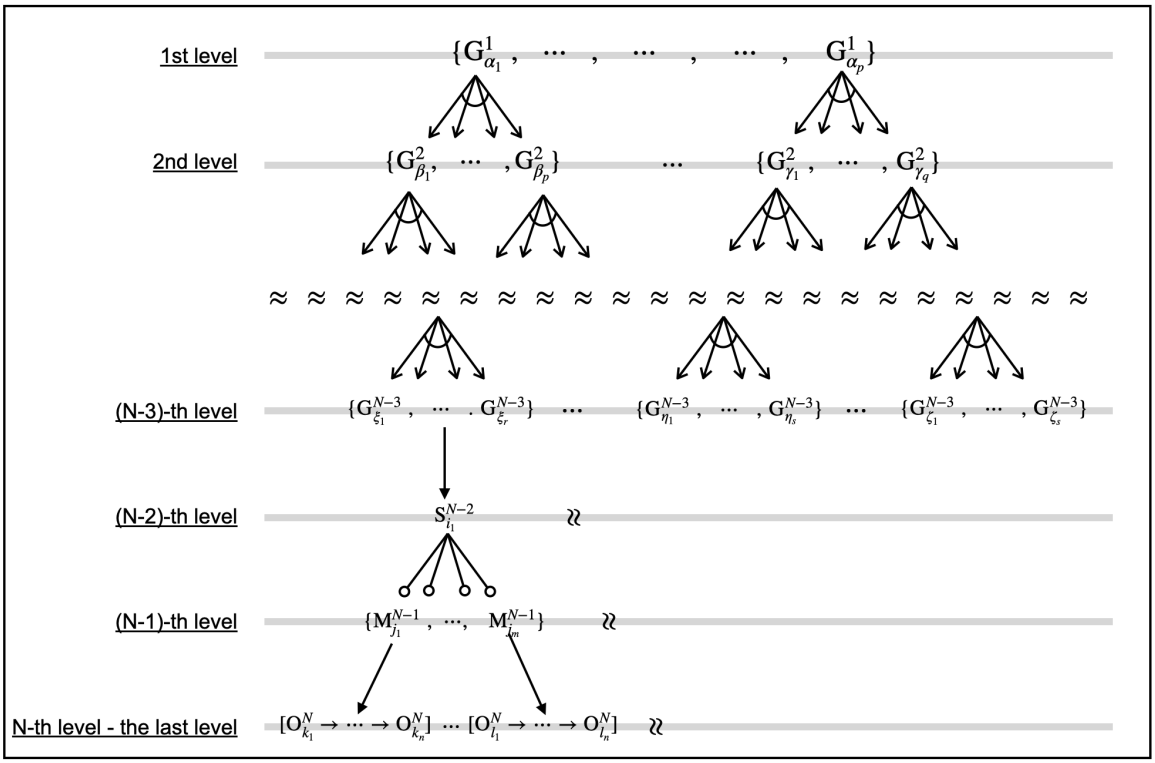


Figure 8. GOMS connection structure

in a general form. The GOMS-construct is explained in the following order starting from the operator sequence, $\mathbf{O} = [\mathrm{O}^N_{k_1} \rightarrow \cdots \rightarrow \mathrm{O}^N_{k_n}]$, which is expanded at the lowest N-th layer and worked upward. Immediately above the $(N-1)$-th layer, the method that connects to this operator sequence, $\mathbf{M} = \mathrm{M}^{N-1}_{j_1}$, exists. This can be regarded as a node that holds a pointer to this operator sequence, $\mathbf{O}$. This method helps connect the goal, $\mathbf{G_i} = \mathrm{G}^{N-3}_{\xi_i}$, which resides two layers above it at the $(N-3)$-th layer, with the operator sequence, $\mathbf{O}$, to achieve it. If there are multiple methods—$\{\mathrm{M}^{N-1}_{j_1}, \cdots, \mathrm{M}^{N-1}_{j_m}\}$—that can achieve $\mathbf{G_i}$, the selection rule, $\mathrm{S}^{N-2}_{i_1}$, is placed between these layers at the $(N-2)$-th layer.

Above the $(N-3)$-th layer, a hierarchical structure of goals is developed. The goals located at the top level, encompassing the first layer $\mathbf{G^1_i} = \mathrm{G}^1_{\alpha_i}$, are expanded into a set of goals, $\mathbf{G^2} = \{\mathrm{G}^2_{\beta_1}, \cdots, \mathrm{G}^2_{\beta_p}\}$, at the second layer. $\mathbf{G^1_i}$ is achieved by meeting all goals contained in $\mathbf{G^2}$. Hereafter, all goals at the first layer are expanded while maintaining this structure until the $(N-3)$-th layer. Note that in Figure 8, for convenience, the top-level goal is placed at the first layer; the N-th layer is represented as the lowest operator layer. However, the depth of the hierarchy varies depending on the content of the top-level goal. Therefore, the concrete value of N varies depending on $\mathbf{G^1_i}$.

The individual G, O, M, S shown in Figure 8 can be considered as nodes that hold pointers connecting them to specific parts of the A-, B-, and C-memes. In real behavioral situations, efficient memory use is required for the smooth operation of the PCM process. Therefore, selecting and executing appropriate actions is important without continuously referring to the Perceptual-Multi-Dimensional Memory Frame, which is activated by P-resonance with the environmental information, while keeping in sync with the environment where the situation changes from moment to moment. Given this, the assumption that "there is an upper bound on the total number of G-, O-, M-, S-nodes available in C-resonance mediated by the GOMS structure" seems reasonable. The total number of goals is denoted as $\hat{G}$, total number of methods as $\hat{M}$, total number of selection rules as $\hat{S}$, total number of operators as $\hat{O}$, average depth of the hierarchy as $\bar{N}$, and upper bound on the number of nodes as $\hat{C}$, which is a constant value. We then consider variations in the overall picture of GOMS-construct created through experience under the condition, $\hat{G} + \hat{O} + \hat{M} + \hat{S} \leq \hat{C}$, based on the relationship between each upper bound.

An operator is an elemental perceptual, motor, or cognitive action; its execution produces a distinguishable change in the actor and/or environment. Since the operator is an elemental part of the construction of the method, $\hat{O}$ is presumed to be much smaller than $\hat{C}$. Then, how are the non-operator available nodes used? The top-level goals are expanded to sub-goals, eventually leading to the determination of a set of achievable methods to be defined under each goal. A method is a kind of goal that can be executed by the operators, pointing to a chunk of the operator sequence connected to it, so that the elements of the set of operators can be used as the material to

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

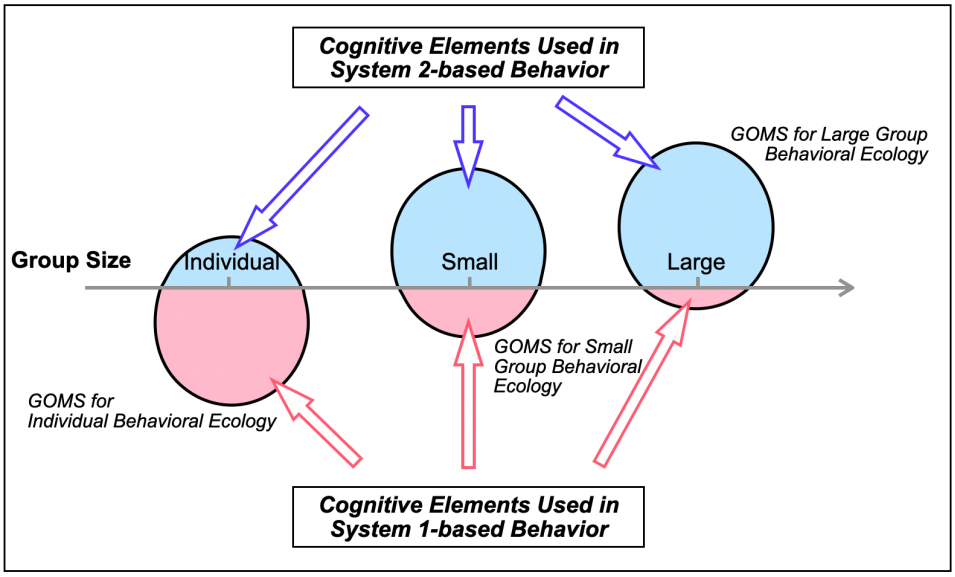


Figure 9. Changes in the Two Minds balance of GOMS components due to differences in behavioral ecology under brain processing capacity (time) constraints

achieve the lowest goal that has been developed. If multiple methods can achieve the goal, one method is selected based on the selection rule that defines the conditions for applying the method.

2) Characteristics of the GOMS-construct Structure: For an event $E(T)$ that occurs at time T, MHP/RT deals with it in its four processing modes [13] outlined below

- In System-2-Before-Event-Mode, MHP/RT consciously considers $E(T)$ beforehand;
- In System-1-Before-Event-Mode, MHP/RT unconsciously adjusts its behavior to the environmental context immediately before $E(T)$. Here, a series of action selections is executed through feedforward processing led by System 1 as a shortcut. During this time, System 2 evaluates the results of the action selections in a timely manner. If it determines that the system is likely to deviate from the expected trajectory or has already deviated, it issues instructions to System 1 for trajectory correction;
- In System-1-After-Event-Mode, MHP/RT unconsciously adjusts the connections within the relevant Perceptual-, Behavior-, and Motor-Multi-Dimensional Memory Frame immediately after $E(T)$; and
- In System-2-After-Event-Mode, MHP/RT consciously reflects on $E(T)$ afterwards to adjust the connections within the Relation- and Word-Multi-Dimensional Memory Frame.

The GOMS-construct that each individual has developed should reflect the results of action selection in System-1-After-Event-Mode and System-2-After-Event-Mode using the A-, B-, and C-memes in the System-2-Before-Event-Mode and System-1-Before-Event-Mode. By allocating more resources—i.e., brain processing power—to System-2-After-Event-Mode, they can construct a richer goal structure. This allows System-2-Before-Event-Mode to devote more resources to making accurate and reliable predictions in various future situations they may encounter. Meanwhile, a sequence of methods involving successively occurring events, $E(T), \cdots, E(T+n)$, can be integrated into a single method by allocating more resources to System-1-After-Event-Mode. The integrated specialized method generates a specialized operator sequence for the corresponding sequence of events. While facing various situations, the number of specialized methods will increase. Due to the limited processing capacity of the brain, either System-2-After-Event-Mode or System-1-After-Event-Mode will become dominant. Therefore, the following is predicted concerning the shape of the GOMS-construct:

- If System-2-After-Event-Mode is dominant, then a goal-rich GOMS-construct, $\hat{G} \gg \hat{M}$, will be constructed.
- If System-1-After-Event-Mode is dominant, then a method-rich GOMS-construct, $\hat{G} \ll \hat{M}$, will be constructed.

3) Relationship between the Number of GOMS Components and Balance of Conscious/Unconscious Processing: Figure 9 shows that the balance between System-1-After-Event-Mode- and System-2-After-Event-Mode-dominance changes depending on the range of communities that the individual is directly and indirectly involved in during their life. In a behavioral ecology where individuals rarely interact with others, each individual can lead a sufficiently problem-free life by having a set of methods that are specific to the situations they encounter. Therefore, the relation, $\hat{M} \gg \hat{G}$, holds. As shown in the left portion of Figure 9, most actions are generated through the unconscious execution of methods by System 1.

In the case of community-based living, each individual is expected to act according to the way they function within the group they belong to. When communication among group members is established in surface language, individuals are unable to perform elaborate inferences. Therefore, the

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

relationship, $\hat{G} > \hat{M}$, is established and is shown in the middle portion of Figure 9.

When a group belongs directly to a community and that community constitutes a society—i.e., the group belongs indirectly to the society—and/or when communication occurs in a structural language, the behavioral ecology becomes System-2-After-Event-Mode-dominant. Further, the relationship, $\hat{G} \gg \hat{M}$, is established, as shown in the right portion of Figure 9. The individuals can flexibly respond to various situations by allocating resources to the execution of System-2-Before-Event-Mode with the careful use of the well-developed goal structure.

Figure 9 also shows the change in the GOMS-construct as the size of the group changes from an individual to a small and then a large group. The number of elements that comprise GOMS, $\hat{C}$, is limited by the constraints imposed on the brain's processing capacities. As the social relationships increase, the number of methods, which are System 1 elements, decreases through the reorganization of the goal structure by abstracting multiple individualized methods together. The elements, which have been used for System 1, are used by the System 2 elements. Meanwhile, the number of System 2 elements increases as the complexity of the relationship increases. That is, by shifting to a behavioral ecology in which System 2 elements are more important than System 1 elements, the composition of elements in the entire GOMS will change to a composition with a rich goal structure that allows for more logical thinking.

B. GOMS and Meme

1) Mutual Development of GOMS and Meme: The existence of memes is a prerequisite for generating GOMS. GOMS also plays an important role in efficient action generation. The generated actions update Multi-Dimensional Memory Frame and contribute to meme formation. Thus, GOMS and memes are in a mutually developing relationship. Actions are generated in two ways: driven by System 1 or System 2. The bias—i.e., the dominance of System 1 or System 2—in the generation of action should affect the aspect of mutual development. This point is discussed in the following.

Figure 5 shows three types of memes. These memes are maintained through generations with imitation as the basic mechanism. A- and B-memes involve physical behaviors that are executed by connecting the Perceptual-, Behavior-, and Motor-Multi-Dimensional Memory Frame. Since A-memes are elemental in generating behavior and B-memes are combinations of elements of A-memes, they are different in granularity and do not mix with each other. The content of the inherited memes is almost invariant, since the content of physical behavior does not significantly change over time. Meanwhile, the C-memes are disconnected from physical behavior. It includes language activities with linguistic symbols and inference by applying rules. The Word- and Relation-Multi-Dimensional Memory Frame are used for these activities. Linguistic symbols and rules are gradually updated under the influence of the social and natural environments surrounding each generation. A-, B-, and C-memes exist in parallel, without mixing with each other, and each is inherited from generation to generation.

GOMS covers orthogonally to the parallel meme structure, and allows A-, B-, and C-meme elements to be combined with each other to efficiently generate effective actions in response to the real-world situations. This is accomplished by combining the elements in Multi-Dimensional Memory Frame in the form of a GOMS-construct. Since words are typical of memes, we can regard memes as carriers of meaning; i.e., *semantics*. GOMS can be thought of as *syntax* because it specifies how words are combined together.

C-memes represent inherited cultures. Cultures are diverse. Based on the discussion in the previous section, Section IV-A3, we can broadly distinguish between cultures that are rich in the goal structure of GOMS, G-culture, and those that are rich in the variety of methods, M-culture. Individuals acting in each culture acquire and act upon the inherited memes of that culture. The memes in G-culture might be updated through System-2-After-Event-Mode, whereas those in M-culture might be updated through System-1-After-Event-Mode. In either case, if the meme is deemed valid within the population in the updated structure, it will trigger a meme update. The update of a meme requires time for validation. Thus, it does not mean that the meme will be immediately updated.

Since a GOMS-construct links goals and operators, it guarantees corporeality for the goals present in the Word- and Relation-Multi-Dimensional Memory Frame. How to achieve a corporeality guarantee is the symbol grounding problem in AI. Figure 6(a) [22] shows that it can be guaranteed in an interpretive process via the Perceptual-Multi-Dimensional Memory Frame. Figure 6(b) shows that the process can be compiled by GOMS to make it more efficient while still guaranteeing corporeality. This ensures that even in G-culture, the development of GOMS for various goals does not dissociate them from the real world. That is, the connection of the Word- and Relation-Multi-Dimensional Memory Frame, to which G belongs, with Behavior- and Motor-Multi-Dimensional Memory Frame, to which M and O belong, guarantees the corporeality of the goal, G. By applying the GOMS-construct to memes, one can make the meme, which is not linked to the real world as it stands, not free from the real world.

2) Common Understanding of Words and GOMS: Words are a typical example of memes [18] and the elements of C-memes. Words are the primary communication medium and are passed on from generation to generation [10]. Individuals make sense of words and understand the situation by referring to the context in which the words have been uttered. However, individual members of a community that share a C-meme may not assign a common meaning to a particular word, even when placed in a common context [21].

Supposedly, the number of words known by native English-speaking adults is 20,000∼30,000, while the number of words used in daily conversation is 3,000∼4,000. Conceptually known words are inherited as the elements of C-memes. However, the words used in daily activities unite the C-meme with the B- and A-memes, which are associated with corporeality, via

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Table I
HAPPINESS GOALS [27] AND THEIR RELATION TO SOCIAL LAYERS. (ADAPTED FROM [21])

Category	No.	Name of Happiness	Types	Social Layers		
				Individual	Family, Relatives and Community	Society
I	8	Painful Happiness	The Masochist	+++		
	11	Tranquil Happiness	The Mediator	+++		
	14	Chemical Happiness	The Drug-taker	+++		
	15	Fantasy Happiness	The Day-dreamer	+++		
II	7	Rhythmic Happiness	The Dancer	+++	+++	
	16	Comic Happiness	The Laugher	+++	+++	
	4	Genetic Happiness	The Relative	+++	+++	
	5	Sensual Happiness	The Hedonist	+++	+++	
III	10	Selective Happiness	The Hysteric	+++	++	
	13	Negative Happiness	The Sufferer	+++	++	
IV	9	Dangerous Happiness	The Risk-taker	+++	++	+
	6	Cerebral Happiness	The Intellectual	+++	+++	++
V	1	Target Happiness	The Achiever	+++	+++	+++
	17	Accidental Happiness	The Fortunate	+++	+++	+++
VI	12	Devout Happiness	The Believer		+++	++
	2	Competitive Happiness	The Winner		+++	+++
	3	Cooperative Happiness	The Helper		+++	+++

+'s denote the degree of relevance of each goal to each layer, i.e., Individual, Community, and Social system, respectively.
+++: most relevant, ++: moderately relevant, and +: weakly relevant.

GOMS. The goals in GOMS represented by symbols belonging to the C-memes are developed into the operators of GOMS belonging to the A-memes. Further, the meaning of goals can be shared as the B-memes as a sequence of operators—i.e., the methods of GOMS—that can be superficially observed as they perform their daily activities.

3) Number of Top Located Goals and Behavioral Ecology: Culture and customs are examples of C-memes. Among them, what is desired to be achieved defines the goal structure of the community that inherits the C-memes. At the top level of the goal structure are the happiness goals [27] that members of a community commonly seek to achieve. Behavioral goals that help achieve the happiness goals exist beneath them [28].

a) Happiness and Behavioral Goals: Happiness and satisfaction felt by humans occur in various forms in behavioral ecology. Many studies have aimed to classify situations when happiness and satisfaction are expressed. Among them, Morris's classification of happiness [27] is consistent with the feeling we attain in our daily life and can be accepted. He conducted comparisons with apes and other primates, and classified the cases wherein many humans shared a sense of happiness in the behavioral ecology that humans represent, which is the closest to the purest state. Based on Morris's classification, a matrix of happiness in contemporary life, organized in relation to the characteristics of the place in which the individual is placed, is shown in Table I. The third, fourth, and second columns indicate the type of happiness indicated by Morris, type of person seeking that happiness, and order in which they are presented in his book, respectively. On the right-hand side, the social fields in which individuals are located are classified into three layers: fields in which only they are involved (individuals), where the presence of others with whom they directly and intimately interact is essential (family, relatives, and community), and where social norms, culture, and language are related to their behavior (social systems). Each happiness goal's relevance to each layer is tentatively assigned and indicated by +'s as follows; +++: most relevant, ++: moderately relevant, and +: weakly relevant. Human beings encounter various situations in the course of their daily lives. Clearly, the 17 happiness goals, are designed to judge whether the behaviors expressed in various situations are good or bad. In addition, certain items are valid for all levels (1. setting and achieving goals and 17. benefits brought about by chance), a certain level (8. pain tolerance, 11. meditation, 14. chemical stimuli, and 15. day dreaming), and multiple levels. Thus, this classification scheme can be used in various human daily activities.

Furthermore, the happiness goals are classified and organized into six categories according to the pattern of the degree of association between individual happiness goals and the social layers. These categories roughly define the context in which the associated behavioral goals are intended to be achieved. The happiness goals in Category I can be achieved by individuals, independent of the community or social system. The happiness goals in Category II are achieved in a non-social context.

Kitajima et al. [29] proposed the maximum satisfaction architecture (MSA). MSA assumes that the human brain pursues one of the 17 happiness goals defined by Morris [27] at every moment and switches to another happiness goal when appropriate by evaluating the current circumstances. Each happiness goal is associated with one or multiple layers of society. Any activity for achieving specific behavioral goals would be conducted by individual persons in the pursuit of any of the 17 happiness goals in the social layers presented in the right portion of Table I. Happiness goals define the person's value structure when they make decisions by running the PCM and memory processes under specific circumstances while selecting their next actions. The achievement of behavioral goals that lead to the attainment of happiness goals utilizes GOMS characterized by the balance shown in Figure 9, depending on the level of the social layer in which the happiness

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

goal is involved. The more uniform the happiness goal, the narrower the range of behavioral goals.

b) Society that Relies on Strong Kinship versus Prevailing Individualism: Recent studies have discussed the characteristics of societies that rely on strong kinship relations and those in which individualism [30]. In Table I, we consider the difference between the two to be the extent to which the happiness goal is realized (the number of cells in the matrix in which the happiness goal is achieved). The latter is considered to be more diverse than the former. Based on our study, we can provide a perspective for understanding the behavioral ecology that forms in these societies. In societies where the C-memes reflecting strong kinship are inherited, the number of goals that can exist is limited. Further, the System 1-driven behavioral ecology is formed as shown in the left portion of Figure 9. This is because the scope of happiness goals is limited and, consequently, that of behavioral goals is also limited. Conversely, in societies with advanced individualism, each individual constructs their own goal structure, thus forming the System 2-driven behavioral ecology as shown in the middle and right portion of Figure 9.

In the latter case, many elements are used to construct the goal structure. Further, flexible action selection is achieved by flexibly replacing the higher-level goals depending on the situation. Since the replacement of the topmost happiness goal also occurs, the behavior is executed by switching between GOMS structures that are quasi-independent of each other. This switching of the GOMS structure can be described as a manifestation of the modalization of behavior, which results in the appearance that an individual switches their behavioral norms depending on the situation.

This also does not necessarily guarantee that even if the same operator sequence is observed, the goal structure developing on top of that sequence is unique. Further, the possibility of misunderstandings—i.e., communication errors—arising from this cannot be excluded. The problems inherent in an individualistic society will appear here.

V. Conclusion

MHP/RT comprises the perceptual, cognitive (Two Minds), and motor processes that synchronously operate with environmental changes, and the Multi-Dimensional Memory Frame that is used during action selection and execution of PCM processes and is updated after action execution. The latter is an internal process of the Multi-Dimensional Memory Frame. Hence, it is executed separately from the PCM process. Meanwhile, for the former, it is necessary to realize the connections between the perceptual process and Perceptual-Multi-Dimensional Memory Frame, and between the cognitive process and Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame. MHP/RT introduces P-resonance for perception and C-resonance for cognition; the connections are realized by *resonance* shown in Figure 4. For P-resonance, we introduced the basic senses as the mechanism in the preceding paper [11].

We examined the mechanism of C-resonance. In the Multi-Dimensional Memory Frame, there are A-, B-, and C-memes, structured by three hierarchies as shown in Figure 5. These memes are mapped to each memory in the Multi-Dimensional Memory Frame and are linked to each other by sharing the Perceptual-Multi-Dimensional Memory Frame as shown in Figure 4. Thus, reality is ensured by perceptual information. Meanwhile, C-resonance works under a time-constrained situation in which the PCM process must select and execute actions while synchronizing with changes in the environment, and connects the Multi-Dimensional Memory Frame and cognitive processes. Here, we introduced GOMS proposed by Card, Moran, and Newell [12] as the mechanism to directly link Word-, Relation-, Behavior-, and Motor-Multi-Dimensional Memory Frame, without going through the Perceptual-Multi-Dimensional Memory Frame.

Each element of GOMS is represented as a node in the brain. The finite number of nodes that can be maintained allows different behavioral ecologies to emerge depending on how the number of nodes allocated to goals and selection rules operated by System 2, and those allocated to methods and operators operated by System 1 are balanced. Regarding the C-memes, we examined the characteristics of behavioral ecology in societies characterized by strong kinship, which inherit simple goal structures, and those with a strong individualistic flavor, which inherit complex goal structures, from the perspective of behavioral goals; these are located under the 17 happiness goals and are developed as a GOMS structure. Although the former is not expected to be flexible in action selection, it can achieve effective and efficient action selection and execution in stable social situations. In the latter, a modalized goal structure is maintained to cope with various situations. The individual flexibly switches to the appropriate goal structure while selecting and executing actions. We also showed that although actions are observed as operator sequences, they are prone to communication errors caused by the non-uniqueness of the goal structure that develops on top of them.

The memes determine the content of the action. The PCM processes determine how to act. Finally, the GOMS structure intersects them. By viewing behavioral ecology from the perspective of GOMS, this study shows that static memes can be implemented and brought to life in behavioral ecology. Behavioral ecology is created by living organisms. Drawing on MHP/RT, the manifestation of memes in behavioral ecology has been clarified in prior case studies [31], [32]. This study proposes a method to give life to static digital information while building on the results of these case studies.

To enable effective and efficient analysis of behavioral ecology across various contexts using the concepts proposed in this paper, guidelines are required for extracting and describing A-, B-, and C-memes, as well as for converting them into GOMS. This mirrors the development of Natural GOMS Language (NGOMSL) [33], which arose from the need to resolve the issue of analysts generating diverse GOMS models. This is one of the issues that must be resolved in the future.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Acknowledgment

This work was supported by JSPS KAKENHI (Grant Numbers 19K12246 / 19K12232 / 20H04290 / 22K12284 / 23K11334) and the National University Management Reform Promotion Project. The authors thank Editage (www.editage.com) for English language editing.

References

[1] M. Kitajima, M. Toyota, J. Dinet, and K. T. Nakahira, "Implementation of Structured Memes into Behavioral Ecology via GOMS", in *COGNITIVE 2025 : The Seventeenth International Conference on Advanced Cognitive Technologies and Applications*, 2025, pp. 6–16.

[2] J. L. McClelland and D. E. Rumelhart, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Psychological and Biological Models*. The MIT Press, Jun. 1986.

[3] D. E. Rumelhart, J. L. McClelland, and P. R. Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. A Bradford Book, Jul. 1987.

[4] A. Damasio, *Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Orlando, FL: Houghton Mifflin Harcourt, 1999.

[5] M. Kitajima and M. Toyota, "Hierarchical Structure of Human Action Selections – An Update of Newell's Time Scale of Human Action", in *Procedia Computer Science, BICA 2014. 5th Annual International Conference on Biologically Inspired Cognitive Architectures*, vol. 41, 2014, pp. 321–325. DOI: http://dx.doi.org/10.1016/j.procs.2014.11.082.

[6] M. Kitajima and M. Toyota, "Guidelines for designing artifacts for the dual-process", *Procedia Computer Science*, vol. 71, pp. 62–67, 2015, 6th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2015, 6-8 November Lyon, France, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.12.201.

[7] L. W. Swanson, *Brain Architecture*. Oxford University Press, 2011.

[8] R. Dawkins, *The Selfish Gene*. New York City: Oxford University Press, 1976.

[9] R. Aunger, Ed., *Darwinizing Culture: The Status of Memetics As a Science*. Oxford Univ Pr on Demand, Jan. 11, 2001, p. 242.

[10] M. Kitajima, M. Toyota, and J. Dinet, "How Resonance Works for Development and Propagation of Memes", *International Journal On Advances in Systems and Measurements*, vol. 14, pp. 148–161, 2021.

[11] M. Kitajima *et al.*, "Basic Senses and Their Implications for Immersive Virtual Reality Design", in *AIVR 2024 : The First International Conference on Artificial Intelligence and Immersive Virtual Reality*, 2024, pp. 31–38.

[12] S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

[13] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)", *Biologically Inspired Cognitive Architectures*, vol. 5, pp. 82–93, 2013, ISSN: 2212-683X. DOI: http://dx.doi.org/10.1016/j.bica.2013.05.003.

[14] M. Kitajima, *Memory and Action Selection in Human-Machine Interaction*. Wiley-ISTE, 2016, ISBN: 9781848219274.

[15] J. R. Anderson and C. Lebiere, *The Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates, 1998.

[16] J. R. Anderson, *How can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press, 2007.

[17] M. Toyota, M. Kitajima, and H. Shimada, "Structured Meme Theory: How is informational inheritance maintained?", in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, B. C. Love, K. McRae, and V. M. Sloutsky, Eds., Austin, TX: Cognitive Science Society, 2008, p. 2288.

[18] D. C. Dennett, *From Bacteria to Bach and Back: The Evolution of Minds*. W W Norton & Co Inc, Feb. 2018.

[19] S. Blackmore, *The Meme Machine*. OUP Oxford, 2000.

[20] F. J. Odling-Smee, K. N. Lala, and M. Feldman, *Niche Construction: The Neglected Process in Evolution: The Neglected Process in Evolution*. Princeton University Press, Feb. 15, 2013.

[21] M. Kitajima *et al.*, "Language and Image in Behavioral Ecology", in *COGNITIVE 2022 : The Fourteenth International Conference on Advanced Cognitive Technologies and Applications*, 2022, pp. 1–10.

[22] M. Kitajima, M. Toyota, and K. T. Nakahira, "Why the Symbol Grounding Problem Matters in Virtual Reality: A Meme-Focused Solution Based on the Model Human Processor with Real-Time Constraints", *International Journal On Advances in Intelligent Systems*, vol. 18, no. 3 & 4, in press, 2025.

[23] H. A. Simon, *The Sciences of the Artificial*, Third. Cambridge, MA: The MIT Press, 1996.

[24] J. R. Anderson, *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

[25] S. Harnad, "The symbol grounding problem", *Physica D: Nonlinear Phenomena*, vol. 42, no. 1, pp. 335–346, 1990, [retrieved: 2, 2025], ISSN: 0167-2789. DOI: https://doi.org/10.1016/0167-2789(90)90087-6.

[26] M. Kitajima, J. Dinet, and M. Toyota, "Multimodal Interactions Viewed as Dual Process on Multi-Dimensional Memory Frames under Weak Synchronization", in *COGNITIVE 2019 : The Eleventh International Conference on Advanced Cognitive Technologies and Applications*, 2019, pp. 44–51.

[27] D. Morris, *The Nature of Happiness*. London: Little Books Ltd., 2006.

[28] M. Kitajima, M. Toyota, and J. Dinet, "Guidelines for Designing Interactions Between Autonomous Artificial Systems and Human Beings to Achieve Sustainable Development Goals", *International Journal On Advances in Intelligent Systems*, vol. 15, pp. 188–200, 2022.

[29] M. Kitajima, H. Shimada, and M. Toyota, "MSA:Maximum Satisfaction Architecture – a basis for designing intelligent autonomous agents on web 2.0", in *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, D. S. McNamara and J. G. Trafton, Eds., Austin, TX: Cognitive Science Society, 2007, p. 1790.

[30] J. P. Henrich, *The weirdest people in the world: How the west became psychologically peculiar and particularly prosperous*. Picador: Farrar, Straus and Giroux, 2021.

[31] M. Kitajima, M. Toyota, and J. Dinet, "Art and Brain with Kazuo Takiguchi - Revealing the Meme Structure from the Process of Creating Traditional Crafts -", in *COGNITIVE 2023 : The Fifteenth International Conference on Advanced Cognitive Technologies and Applications*, 2023, pp. 1–10.

[32] K. T. Nakahira, M. Kitajima, and M. Toyota, "Understanding Practice Stages for a Proficient Piano Player to Complete a Piece: Focusing on the Interplay Between Conscious and Unconscious Processes", *International Journal on Advances in Life Sciences*, vol. 16, pp. 164–177, 2024.

[33] D. Kieras, "GOMS modeling of user interfaces using NGOMSL", in *Conference Companion on Human Factors in Computing Systems*, ser. CHI '94, Boston, Massachusetts, USA: Association for Computing Machinery, 1994, pp. 371–372, ISBN: 0897916514. DOI: 10.1145/259963.260467.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Productive Simplification: Complexity Production by Multiple Reductive Simplification Procedures for Interpreting Multi-Layered Neural Networks

Ryotaro Kamimura
Tokai University
Kitakaname, Hiratsuka, Kanagawa, Japan
e-mail: ryotarokami@gmail.com

***Abstract*—The present paper aims to show that there exists a strong simplification force inside neural networks. However, it is naturally necessary to introduce some complexity into connection weights to make learning possible. This means that simplification must be violated to achieve appropriate learning. To address this issue, we assume that the simplification principle is always maintained, and the only way it can be (seemingly) violated is by combining several different types of simplification procedures. This combination, which adds some complexity, can be called "productive simplification," meaning that the added complexity arises as a result of the interaction among many different simplification procedures. The simplification principle still holds, but it can be modulated by various types of these very simplification processes. To demonstrate productive simplification, we use two types of simplification: local and distributed simplification. Local simplification aims to reduce the number of components, while distributed simplification attempts to simplify the roles of each component as much as possible. These two types of simplification are unified in productive simplification, and in practice, some complexity can thereby be introduced. Productive simplification was applied to an artificial dataset that incorporated both linear and non-linear relations. The results show that the unification of local and distributed simplification appeared in the form of a phase transition, where generalization performance changed significantly in this transition phase. This phase transition demonstrates the existence of a productive unification of different types of simplification, ultimately strengthening the simplification force inside neural networks.**

Keywords-simplification; productive; reductive; local; distributed; diachronic

I. Introduction

This paper aims to demonstrate that the simplification force should be observed at all levels of neural learning, even in situations where some complexity in connection weights is required [1]. This apparent contradiction can be resolved by assuming that different types of simplification forces are combined to produce the effect of complexity in connection weights, though only seemingly. This means that complexity is actually the result of many simplification forces operating inside neural networks.

A. Simplification Principle

The present paper aims to show that there exists a strong simplification force inside neural networks. These simplification forces should be observed internally and in many different ways. Internal simplification lies in the existence of self-organizing forces that simplify internal network configurations as much as possible, and this can be realized even without information from external inputs. Naturally, this is too idealized, and in practice it must take input information into account. Thus, it becomes necessary to add some complexity to connection weights during learning, which seems contradictory to the hypothesis of simplicity. This contradiction can be resolved by supposing that simplification can take many different forms. Although these forms may appear contradictory to each other, their interaction can be used to introduce complexity into the connection weights. This means that complexity is in fact the outcome of the simplification principle, which operates universally in learning.

B. Reductive and Productive Simplification

As mentioned above, it is often necessary to introduce some complexity into connection weights to make learning possible. This implies that simplification must be weakened or seemingly violated for appropriate learning. To address this issue, we assume that the simplification principle always holds, and that the only way it can be (seemingly) violated is by combining several different simplification procedures. This combination, which adds some complexity, can be called "productive simplification," meaning that complexity arises as a consequence of the coexistence of many different simplification procedures. In other words, the simplification principle continues to operate, but it can be modulated by employing different types of simplification.

To emphasize the importance of productive simplification, we distinguish between two types of simplification: reductive and productive. Reductive simplification refers to the usual meaning of simplification: reducing the complexity of network configurations as much as possible, as many previous learning procedures for improving generalization and interpretation have done. In contrast, productive simplification aims to add some complexity into network configurations while still adhering to the simplification principle. This means that in each computational step, the simplification principle must be strictly followed, but the combination of these procedures can result in the addition of complexity necessary for learning. In this sense, productive simplification shows that even when complexity is introduced, the simplification principle continues to prevail.

C. Local and Distributed Simplification

The general simplification principle governs learning processes, but in practice, many different types of simplification forces are likely to be present. This arises from attempts

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

to achieve simplified configurations from as many different perspectives as possible. Among these, we consider two types of simplification forces as a first approximation in this paper: local and distributed simplification. In general, simplification is assumed to reduce the total strength of connection weights. Local simplification seeks to reduce the number of strong connection weights as much as possible. In this case, only one weight retains significant strength, while the others are forced to become very small. This ultimately reduces the total strength of the connection weights, with only one weight playing an essential role in information processing. On the other hand, distributed simplification seeks to reduce the strength of all connection weights as much as possible. In this case, all weights retain small, evenly distributed strength values. When a small number of weights are known to play important roles, local simplification is appropriate. Conversely, when no prior knowledge is available about which weights are more important, distributed simplification should be used. In productive simplification, these two types of forces are combined, producing an effect that weakens the total simplification force. It should be noted that even this weaker form of simplification still arises from the combination of two distinct simplification strategies.

D. Main Contributions

The main contributions of this paper can be summarized as follows:

- The present paper aims to clarify the simplification forces hidden in multi-layered neural networks.
- Even in the process of adding complexity during learning, the simplification principle continues to operate internally. This can be realized through productive simplification, which combines multiple reductive simplification procedures.
- We propose two types of simplification forces: local and distributed simplification. These two types should be unified, and through their unification, some complexity can be introduced while still adhering to the simplification principle.
- We conducted experiments to demonstrate how this unification of simplification types occurs. The results show that, initially, local simplification dominates, and then a phase transition occurs to unify local and distributed simplification.
- The results indicate that simplification forces exist in neural networks. The main mechanism lies in strengthening simplification through the unification of different simplification forces. This suggests that behind seemingly complexity-adding operations, strong simplification forces are always at work.

E. Paper Outline and Organization

In Section II, we show that our method of productive simplification can be regarded as one of the first attempts to transform conventional reductive simplification procedures into a productive and unified approach. In Section III, we introduce local and distributed potentiality, which are consumed to achieve simplicity. The unification of these two types of potentiality is realized through diachronic potentiality, where unification is controlled in a time-dependent manner. In addition, we provide a brief explanation of computational methods such as compression and ratio potentiality for comparison. In Section IV, we present experimental results obtained using productive potentiality. We first show that simplification can be modulated by introducing productive potentiality control. Then, during learning, a kind of phase transition occurs, demonstrating a compromise between local and distributed simplification. Finally, simplification tends to make networks as linear as possible despite the presence of many nonlinear properties in neural networks. In Section V, we summarize the paper and suggest future work, such as exploring the relationship between productive simplification and the problem of simplicity versus complexity in neuroscience.

II. Related Work

We have stated the existence of multiple types of simplification procedures in the complexity addition processes of productive learning. Almost all conventional methods, however, seem to be confined to reductive simplification. Here, we describe four conventional reductive simplification procedures, emphasizing that all of them aim to address reductive simplicity without sufficient consideration of the productive property. The reductive procedures must be complemented by productive simplification, whereby certain complexities are reintroduced as a result of the simplification itself.

A. Prototype Simplification

The prototype approach has been one of the major simplification procedures in neural networks. Here, we use the prototype to describe the simplest form achievable through the simplification forces inherent in the network. Prototype-based approaches have been studied extensively since the early stages of learning, under names such as vector quantization, competitive learning, and self-organizing maps [2]–[7]. These methods aim to identify a small number of representative vectors for many input patterns. In recent deep learning research, prototype learning has gained renewed attention [8], as the volume of input patterns to be processed has grown larger and increasingly heterogeneous. Simplifying these complex and heterogeneous patterns has become urgent, giving rise to methods like one-shot and few-shot learning, which represent many inputs using a few representative patterns. Similarly, zero-shot learning [9] has been developed to incorporate the abstract and semantic properties of input patterns. These methods exemplify a typical reductive simplification procedure, representing a large number of input patterns with a smaller set of representative inputs and more abstract features. In contrast, our approach focuses on the network itself, aiming to determine which network configuration should be organized when all network components are given before learning. Furthermore, we attempt to unify multiple types of simplification to introduce complexity.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

B. Interpretable Simplification

Second, most interpretation methods can be classified as reductive simplification. Among these, a popular approach is localized interpretation, which focuses on specific instances rather than the overall inference mechanism. Unlike global interpretation methods [10]–[12], localized interpretation has proven effective in practical applications due to its simplicity and the urgent demand for explainability. Linear and local simplifications have been widely used in interpretation methods, replacing complex non-linear models with corresponding linear models [13]–[15]. These simplified local models aim to interpret surface networks, which are diverse and heterogeneous. However, because the main characteristics of surface networks lie in their diversity, such diversity hinders interpretation. It is therefore necessary to seek a simpler and unified prototype for interpretation. Our approach seeks to uncover the prototype hidden deep within surface models. Surface models are produced through transformational rules and may be too complex to understand, but the underlying prototype is expected to be much easier to interpret.

C. Distilling Simplification

Third, network compression simplifies neural networks by replacing complex networks with smaller ones [16]–[20]. This type of simplification has become increasingly necessary as networks grow larger to process more input patterns. However, this approach does not allow us to interpret how all components in such networks operate to produce outputs. Most compression methods represent typical external and reductive simplification, replacing the original network with smaller and often unrelated networks. These methods attempt to interpret the smaller "student" networks under the assumption that their inference mechanism is similar to the original network's. However, this assumption may not always hold true. In contrast, the productive simplification we propose compresses the original multi-layered neural network directly, retaining the original network's information. Thus, interpreting and explaining the compressed networks is more directly connected to the original multi-layered networks. Conventional compression is not productive in the sense that it cannot directly relate the original networks to compressed ones for interpretation.

D. Mutual Information-Theoretic Simplification

Fourth, information-theoretic methods also relate to network simplification, though in more abstract ways. Multiple network configurations can be represented by simpler and more abstract information content, and the objective of learning can be considered as necessary information acquisition. In particular, mutual information has played a significant role in neural networks. Well-known examples include maximum mutual information preservation [21]–[24], which aims to retain as much relevant information as possible, and the information bottleneck method [25]–[28], which seeks to maximize necessary information while minimizing unnecessary information.

Despite their utility, mutual information-based methods face difficulties, such as computational complexity and ambiguous interpretations of the results. These issues arise because mutual information involves contradictory operations: entropy maximization and conditional entropy minimization. Recent information bottleneck methods introduce additional mutual information computations to balance compression and relevant information preservation [29], but computational complexity remains a challenge. The main procedure is to reduce information on inputs while retaining appropriate information about targets. This is similar to distributed simplification.

Mutual information maximization can, in fact, be viewed as addressing both local and distributed simplification, as is the case with the productive simplification in this paper. We believe it can be applied to the problem of complexity and simplicity in neuroscience [30]. However, due to the computational complexity inherent to mutual information, its productive property of simplification has not yet been realized [24][31][32]. The present attempt at productive simplification aims to take the first step toward realizing the actual potential of mutual information in neural networks.

III. Theory and Computational Methods

Here, we introduce how to compute two types of simplification: local and distributed. Then, we briefly explain how to combine these simplification forces to add complexity. In addition, we give a short explanation for better understanding of the experimental results and how to compress a neural network into the corresponding prototype. Since our method is an extension of entropy, we also compare it with conventional entropic methods, showing how our approach can enhance learning characteristics in a simplified manner.

A. Theoretical Procedures

1) Simplification Principle: The present paper aims to show how to use the simplification principle to add some complexity to network configurations, which is necessary for realizing appropriate learning processes. Figure 1 shows a process of productive simplification, transforming a network from a surface network into a simpler one. For the surface network, two types of reductive simplification are applied. Local simplification aims to reduce the number of important connection weights as much as possible, whereas distributed simplification aims to reduce the strength of all connection weights. Both procedures are unified in productive simplification, producing an intermediate multi-layered neural network. In addition, this productive simplification is adjusted in a diachronic manner, meaning that the ratio of local to distributed simplification is controlled throughout the course of learning. Finally, the intermediate network is compressed into the simplest form, namely, a prototype network.

2) Simplification Potentiality Consumption: First, we introduce the basic simplification indexes used in this paper. The overall measure of simplification is defined as the strength of absolute connection weights. When this simplification potentiality, or the strength of absolute weights, becomes smaller, the network becomes simpler by definition. For simplicity, we consider only one layer, from the nth layer to the $n+1$th layer,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

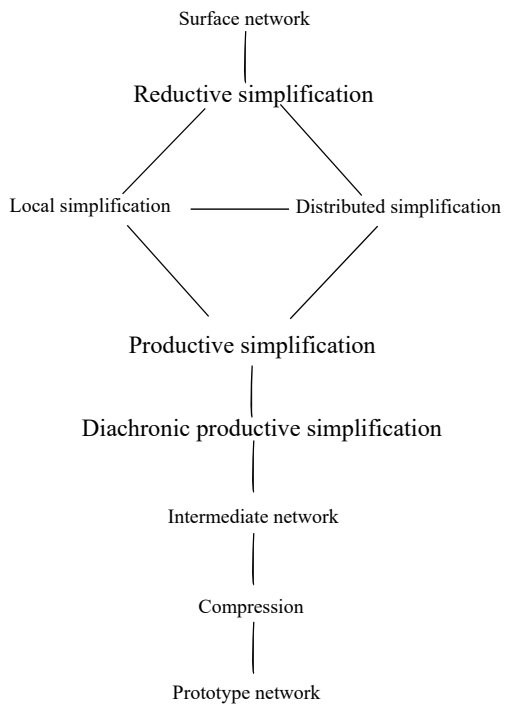


Figure 1. Diachronic productive simplification from a surface network to a prototype network by unifying local and distributed reductive simplification.

denoted as $(n, n+1)$ in Figure 2. The individual simplification potentiality for $(n, n+1)$ is computed as:

$$u_{jk}^{(n)} = \mid w_{jk}^{(n)} \mid, \tag{1}$$

where the notation for the layer is simplified from $(n, n+1)$ to (n), and all potentialities (absolute weights) are assumed to be greater than zero. The simplification potentiality for one layer is obtained by summing all individual simplification potentialities, and by summing across all layers, we obtain the total potentiality:

$$U = \sum_{njk} u_{jk}^{(n)}. \tag{2}$$

As this potentiality becomes larger, the network has a higher capacity to be simplified. The simplification potentiality should be reduced or consumed to achieve the simplest network.

3) Local Potentiality Consumption: As mentioned above, simplification potentiality must generally be reduced or consumed. However, there are several possible ways to achieve

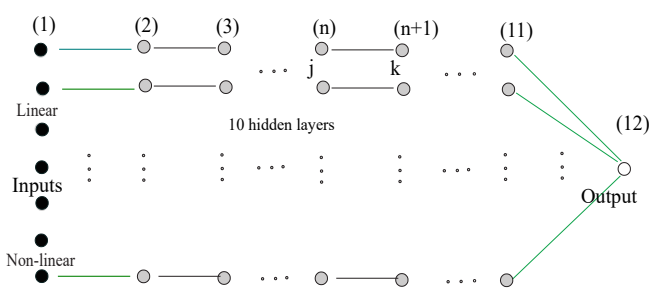


Figure 2. A network architecture with ten hidden layers.

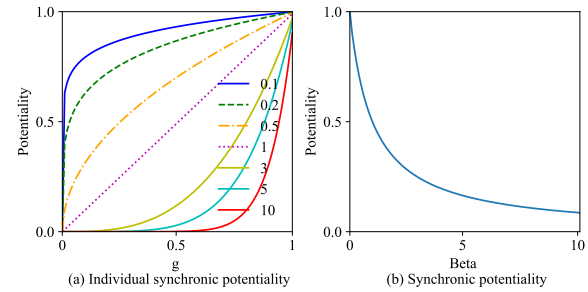


Figure 3. Individual local potentiality (a) and local potentiality (b) as a function of the parameter β_{loc}. All potentialities are normalized for comparison.

this reduction. One popular method can be called "local simplification potentiality." Local simplification potentiality, or simply local potentiality, aims to reduce the strength of all connection weights except a small number of specific ones. Because the majority of weights become smaller, the total simplification potentiality decreases, even though a small number of stronger weights remain.

Now, by normalizing the individual potentiality by its maximum value, we define the relative local potentiality as:

$$g_{jk}^{(n)} = \left[\frac{u_{jk}^{(n)}}{\max_{j'k'} u_{j'k'}^{(n)}} \right]^{\beta_{loc}}, \tag{3}$$

where β_{loc} is a parameter controlling the strength of local potentiality. In particular, when $\beta_{loc} = 1$, the potentiality corresponds to the base case. By summing across all layers, we obtain the local potentiality:

$$G = \sum_{njk} \left[\frac{u_{jk}^{(n)}}{\max_{j'k'} u_{j'k'}^{(n)}} \right]^{\beta_{loc}}. \tag{4}$$

Figure 3(a) illustrates the individual local potentiality for different parameter values. As the parameter β_{loc} increases, the majority of connection weights are forced to decrease, thereby reducing the potentiality. Figure 3(b) shows the local potentiality as a function of β_{loc}. As explained above, the local potentiality decreases gradually as the parameter increases.

Using this local potentiality, the new weights at the $(t+1)$th learning step are obtained by multiplying the weights at the tth step by the corresponding potentiality:

$$w_{jk}^{(n)}(t+1) = g_{jk}^{(n)} w_{jk}^{(n)}(t). \tag{5}$$

4) Distributed Potentiality Consumption: The distributed potentiality is the inverse of the local potentiality and aims to make the strength of all connection weights as small as possible. By normalizing the inverse of individual potentiality by the maximum value, we can define the relative distributed potentiality as:

$$h_{jk}^{(n)} = \left[\frac{\max_{j'k'} u_{j'k'}^{(n)} - u_{jk}^{(n)}}{\max_{j'k'} u_{j'k'}^{(n)}} \right]^{\beta_{dis}}, \tag{6}$$

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

where β_{dis} is a parameter controlling the strength of the potentiality. By summing all the relative potentialities, we obtain the final form of distributed potentiality:

$$H = \sum_{njk} \left[\frac{\max_{j'k'} u^{(n)}_{j'k'} - u^{(n)}_{jk}}{\max_{j'k'} u^{(n)}_{j'k'}} \right]^{\beta_{dis}}. \tag{7}$$

Using this distributed potentiality, new weights at the $(t+1)$th learning step are obtained by multiplying the weights at the tth step by the corresponding potentiality:

$$w^{(n)}_{jk}(t+1) = h^{(n)}_{jk} w^{(n)}_{jk}(t). \tag{8}$$

Finally, we note that the distributed potentiality is simply the inverse of local potentiality for the base case ($\beta = 1$). In this case, it can be obtained by taking the inverse of the local potentiality.

5) Productive Potentiality Consumption: After defining the local and distributed potentiality, we need to combine these two types to produce complexity, namely, productive potentiality control. We combine the two types of potentiality as:

$$P = \alpha G + \bar{\alpha} H, \tag{9}$$

where the parameter α ranges between zero and one and is introduced to control the ratio of local to distributed potentiality. Note that $\bar{\alpha}$ corresponds to $1 - \alpha$. The unification of local and distributed simplification is simply the sum of the two potentialities when treated synchronically. However, the process of unification must be controlled more subtly to achieve better generalization with simplification, since generalization is considered the most important factor in neural networks.

6) Diachronic Productive Potentiality Consumption: This unification can be realized by defining a new type of potentiality called "diachronic". The diachronic potentiality considers the entire sequence of learning steps. The individual diachronic potentiality at the tth learning step (epoch) is defined solely based on the time step t:

$$v_t = t. \tag{10}$$

The individual diachronic potentiality is then defined as

$$z_t = \left[\frac{v_t}{\max_{t'} v_{t'}} \right]^{\beta_{dch}}. \tag{11}$$

We invert this original diachronic potentiality to obtain a distributed type of simplification. The new individual diachronic potentiality is obtained by inverting the original:

$$\bar{z}_t = 1 - z_t. \tag{12}$$

This equation is used to ensure that the sum of the two types of potentiality equals one. In addition, the local potentiality is used at the beginning of learning, followed by the introduction of distributed potentiality. The diachronic potentiality is then used to combine the local and distributed potentiality as

$$w^{(n)}_{jk}(t+1) = \left(\beta_t g^{(n)}_{jk} + \bar{\beta}_t h^{(n)}_{jk} \right) w^{(n)}_{jk}(t), \tag{13}$$

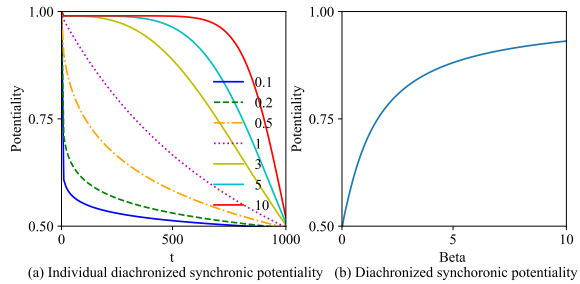


Figure 4. Individual diachronic potentiality (a) and diachronic potentiality (b) as a function of the parameter β_{dia}.

where $\beta = \bar{z}$ and $\bar{\beta} = z$. The effect of diachronic potentiality increases as learning progresses. As the diachronic parameter β_{dch} increases, the effect of local potentiality becomes increasingly mixed with the effect of distributed potentiality.

To explain how diachronic potentiality controls local and distributed potentiality, Figure 4(a) illustrates the diachronic potentiality for different parameter values. As the parameter β_{dia} increases, the majority of individual potentialities increase. This means that the local potentiality becomes more effective throughout learning as the parameter increases. Figure 4(b) shows the inverse of diachronic potentiality as a function of β_{dia}. As explained, the inverse diachronic potentiality increases gradually with increasing parameter strength. By using diachronic potentiality, local potentiality is applied first, followed by the introduction of distributed potentiality. As the parameter increases, the local potentiality remains effective for a longer period. This sequential operation improves the generalization performance of the final neural network.

B. Computational Procedures

1) Compression: To obtain an estimated prototype, we must compress the original multi-layered neural network into the corresponding simplest network without hidden layers. This assumes that all activation functions are linear, although they are actually non-linear. Our simplicity principle aims to make activation functions as linear as possible. If the compressed networks differ from actual linear networks, these differences can be used to specify the effects of non-linearity in neural networks.

Suppose we attempt to compress a twelve-layered neural network, including the input and output layers in Figure 2, into a network without hidden layers. The layers are numbered from No. 1 (input) to No. 12 (output), with ten hidden layers in between. This configuration was used in our experiments, described in the next section. We begin the first compression by:

$$w^{(1,3)}_{ik} = \sum_j w^{(1,2)}_{ij} w^{(2,3)}_{jk}, \tag{14}$$

where layer (1,2) is combined with layer (2,3) to produce the compressed layer (1,3). Repeating this process, we obtain the compressed weights connecting the first and eleventh layers,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

denoted $w_{im}^{(1,11)}$. Using these weights, we finally obtain the fully compressed weights for (1,12):

$$w_i^{(1,12)} = \sum_m w_{im}^{(1,11)} w_m^{(11,12)}, \tag{15}$$

where (1,12) indicates that the twelve-layered network is compressed into a two-layered one (1,12).

This compression method differs from conventional approaches [33][34] because it aims to interpret internal representations as much as possible. Our method seeks to understand the inference mechanism of neural networks while preserving the original information in a simplified form.

2) Ratio Potentiality: In addition to the standard potentialities explained above, we introduce the ratio potentiality for comparing estimated and supposed prototypes. Using a twelve-layered network for simplicity, the ratio of compressed individual potentiality of the compressed network to the supposed individual potentiality of the prototype network can be computed by:

$$r_i^{(1,12)} = \frac{g_i^{(1,12)}}{g_i^{(1,2)}}. \tag{16}$$

Here, (1,12) and (1,2) denote the estimated and supposed networks, respectively. For this ratio r_i, we compute the ratio potentiality:

$$v_i^{(1,12)} = \frac{r_i^{(1,12)}}{\max_{i'} r_{i'}^{(1,12)}}. \tag{17}$$

Then,

$$V^{(1,12)} = \sum_i v_i^{(1,12)}. \tag{18}$$

When all ratio potentialities are equal, the ratio potentiality is larger. If only one ratio potentiality is larger than the others, the ratio potentiality is smaller. This metric can be used to assess the similarity between compressed and prototype networks.

3) Entropy and Divergence: Potentiality has been introduced to simplify the entropy function, and it is necessary to compare potentiality with the corresponding entropy. Our entropy and related divergence are defined as follows. The relative potentiality for the supposed prototype network is computed as

$$q_i^{(1,2)} = \frac{u_i^{(1,2)}}{\sum_{i'} u_{i'}^{(1,2)}}. \tag{19}$$

Next, the relative potentiality for the compressed network is

$$p_i^{(1,12)} = \frac{u_i^{(1,12)}}{\sum_{i'} u_{i'}^{(1,12)}}. \tag{20}$$

Entropy is defined as

$$S^{(1,12)} = -\sum_i p_i^{(1,12)} \log\ p_i^{(1,12)}. \tag{21}$$

Entropy decreases when only one weight becomes large while the others are small. Conversely, when entropy is large, all weights are equal. This behavior is similar to the property of potentiality used in this paper. Divergence, as a complementary measure, is computed as

$$D^{(1,12)} = \sum_i p_i^{(1,12)} \log\ \frac{p_i^{(1,12)}}{q_i^{(1,2)}}. \tag{22}$$

Divergence corresponds to ratio potentiality and decreases as the potentialities of the two networks become more similar.

IV. Results and Discussion

The experimental results show that two types of simplification can be unified to produce some complexity. Then, this productive simplification can produce a kind of phase transition in connection weights. Through this transition, the final networks were forced to approach the supposed prototype. This phase transition can also improve generalization, meaning that simplification should be considered a necessary step for improving generalization and interpretation. However, even in cases with better generalization and added complexity, simplification must still be operating inside neural networks.

A. Experiment Outline

The data set was created by imitating an actual business data set used to estimate the bankruptcy of companies [35]. The objective of the experiment is not to improve generalization but to interpret how bankruptcy occurs and what the major causes of bankruptcy are. In more technical terms, we aim to understand relationships between inputs and outputs. In the actual data set, linear and non-linear relationships are naturally mixed, making it seemingly impossible to explicitly understand the relationships between inputs and outputs. To address this issue, we created a data set with both linear and non-linear relationships between inputs and targets artificially. This artificial data set allows us to explicitly examine how neural networks respond to specific inputs to produce outputs. The number of input variables was seven. Of these, input No. 5, input No. 6, and input No. 7 were created non-linearly using exponential and sine functions, while the remaining inputs were linear.

The number of input patterns was 1000, and the number of hidden layers was ten, with ten neurons in each hidden layer. We used the PyTorch program package, with almost all parameter values set to default to ensure easy reproduction of the results presented here. The experiment was designed to make the neural networks as close as possible to the prototype network, which is assumed to be hidden within surface networks. The prototype network is the simplest form, and connection weights in this paper were computed using correlation coefficients between inputs and targets of the training data set.

Figure 5 shows a supposed prototype network computed by correlation coefficients between inputs and outputs from the artificial data set. As shown in the figure, the first four inputs are strongly correlated with outputs, whereas the last three inputs (No. 5 to No. 7) were only weakly correlated with outputs. As already mentioned, these three inputs were created non-linearly, and the linear correlation coefficients could not

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

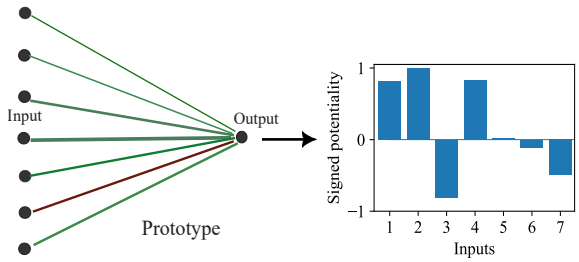


Figure 5. Supposed prototype (left) and the normalized strength of connection weights of the prototype network (right), computed by correlation coefficients between inputs and targets.

capture those relations. The final multi-layered neural network was then compressed into the simplest network without hidden layers. Compression was conducted layer by layer. We then compared the estimated and compressed networks with the assumed simplest prototype network.

A set of potentiality values was used to improve generalization performance, since generalization was considered the most important factor for evaluating network performance. However, we should note that the purpose of this study is to interpret how networks self-organize to achieve the simplest configuration. The parameters β_{loc} and β_{dis} were set to 0.005 to stabilize the learning process. The parameter β_{dch} was set to 2, and we first used local potentiality, gradually introducing distributed potentiality.

Before explaining local and distributed potentiality, it should be noted that they were computed only in the base case for illustrating the figures. In this case, distributed potentiality is actually the inverse of local potentiality, and therefore only the results of local potentiality are shown.

Now, the main findings can be summarized as follows:

- Productive simplification could control local and distributed potentiality forces to produce networks with better generalization.
- Connection weights with better generalization were obtained through a kind of phase transition during learning.
- Those connections with better generalization showed clearer characteristics even in hidden layers.
- By controlling local and distributed potentiality, we could produce the final and estimated prototypes, close to the supposed prototype.
- The experimental results show that complexity added during learning should be based on simplification procedures.

B. Potentiality and Entropy

The method could decrease the strength of total simplification potentiality. In addition, productive potentiality could control the strength of total and local potentiality appropriately for better generalization.

Figure 6 shows the total simplification potentiality (left), local potentiality (middle), and entropy (right). When only local potentiality was used in Figure 6(a), simplification potentiality (left) decreased gradually, and local potentiality (middle) decreased sharply. Similarly, entropy (right) decreased rapidly as the number of learning steps increased. When the diachronic parameter was two, corresponding to the best generalization performance in Figure 6(b), the total simplification potentiality was small but remained nearly constant throughout learning. Local potentiality decreased in the first stage and then increased gradually. This indicates that the effect of distributed potentiality appeared in the later stage of learning. Entropy showed the same trend, though less clearly. Productive simplification was able to control local potentiality while keeping total potentiality nearly unchanged. Figure 6(c) shows the results when distributed potentiality was used. Simplification potentiality decreased very slowly, while local potentiality increased gradually. However, entropy did not show a clear trend. Finally, Figure 6(d) shows the results with the conventional method without potentiality control. Simplification potentiality increased considerably, while local potentiality decreased only slightly, and entropy again did not show a clear trend.

The results show that productive simplification control can adjust the strength of both total simplification potentiality and local potentiality appropriately. Local potentiality can be decreased and then increased to introduce complexity. On the other hand, entropy, although conceptually related to potentiality, could not clearly represent these properties. Our potentiality method, while close to entropy, can more explicitly capture the characteristics of connection weights due to its computational simplicity.

C. Weights for All Layers

Better generalization was obtained through productive potentiality control and by making full use of all hidden layers with explicit characteristics.

Figure 7 shows the weights of all layers for the four methods. When only the local potentiality was used in Figure 7(a), a stronger weight could be seen in most layers. The local potentiality attempted to achieve simplification by weakening all connection weights except for a small number of specific ones. However, in the middle layers, these stronger weights became smaller, indicating that important weights could not be detected there. This means that under local potentiality, hidden layers in the middle could not acquire important information, and only the layers closer to the input and output played important roles in learning. When the diachronic parameter was set to two, yielding the best generalization in Figure 7(b), weights were arranged symmetrically. This indicates that important and non-important weights were explicitly separated, and this tendency was observed even in the middle layers. Better generalization was thus obtained by fully utilizing all hidden layers. When only the distributed potentiality was used, the symmetric arrangement of weights became somewhat blurred, as shown in Figure 7(c). Finally, when the conventional method without any potentialities was applied in Figure 7(d), all weights became obscure, showing almost no regularity.

The results show that better generalization was obtained by using all hidden layers and by explicitly detecting some

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

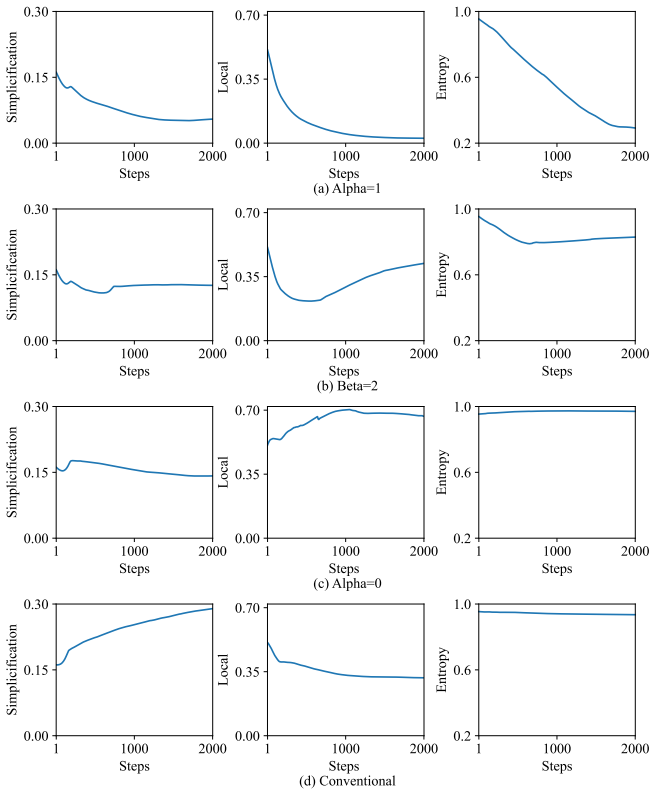


Figure 6. Total simplicity (left), local (middle), and entropy (right) as a function of the number of learning steps (epochs). Local potentiality $\alpha = 1$ (a), diachronic potentiality ($\beta = 2$) with best generalization (b), distributed potentiality $\alpha = 0$ (c), and conventional method without potentiality control (d).

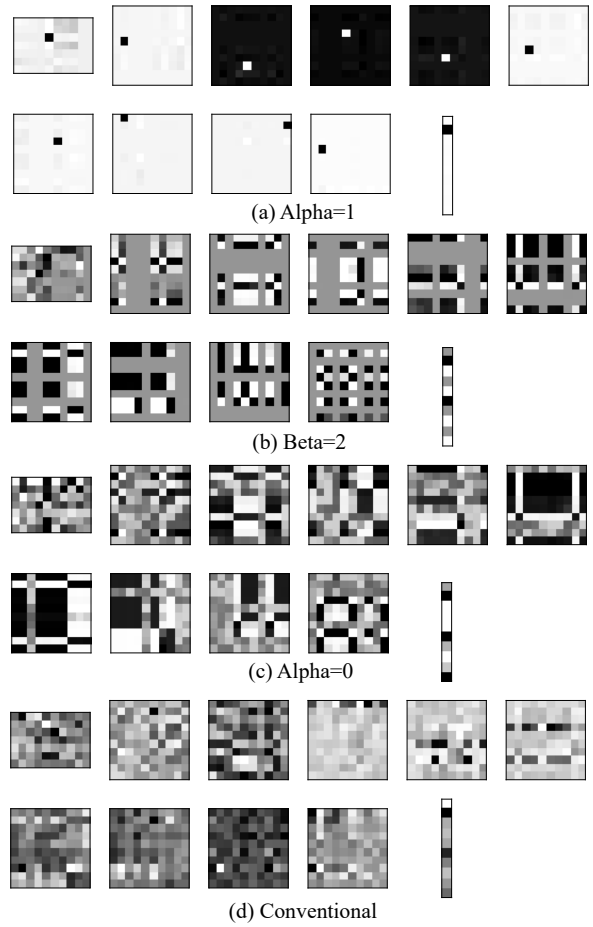


Figure 7. Weights for all layers, when the local potentiality was used (a), the diachronic potentiality was used (b), the distributed potentiality was used (c), and the conventional method was used (d).

characteristics in those layers. When such characteristics could not be detected, better generalization was not obtained.

D. Layer Potentiality

The layer potentiality demonstrated that better generalization could be obtained by considering hidden layers in the middle in addition to the input and output layers. These hidden layers seemed to resemble those obtained with the conventional method. The layer potentiality was computed by summing the individual local potentiality for each layer.

Figure 8(a) shows the results when only the local potentiality was used for learning. As can be seen in the figure, the input layer initially had the largest potentiality value, and the potentialities of subsequent layers gradually decreased, with the output layer having the largest potentiality at the end. Figure 8(b) shows the results when the diachronic parameter was set to two, corresponding to the best generalization. Compared with the results from using local potentiality alone, the sixth layer's potentiality increased in the later stage of learning. This indicates that hidden layers attempted to detect certain characteristics for better generalization. When only the distributed potentiality was used, as shown in Figure 8(c), layer No. 7 had a larger potentiality. Note, however, that better generalization under productive potentiality considered layer No. 6 as important. Finally, when the conventional method was applied, as in Figure 8(d), layer No. 6 became larger.

The results show that productive potentiality could utilize hidden layers in the middle. The characteristics detected by these layers appeared to be similar to those identified by the conventional method. The productive method attempted to enhance the characteristics found by the conventional method.

E. Ratio Potentiality

The results confirmed that productive potentiality aimed to increase the ratio potentiality as much as possible. This produced a kind of phase transition, leading to structural changes in connection weights that improved generalization. The ratio potentiality represents the ratio of the individual potentiality of a compressed network to that of the supposed prototype, computed using the correlation coefficients of the data set. When ratio potentiality increases, the similarity between the estimated and supposed prototype increases. The divergence between compressed and supposed prototypes was also computed for comparison. A smaller divergence indicates greater similarity.

Figure 9 shows the ratio potentiality (left), divergence (middle), and generalization performance (right). When only the local potentiality was used, as in Figure 9(a), the ratio potentiality was initially high but gradually decreased. This

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

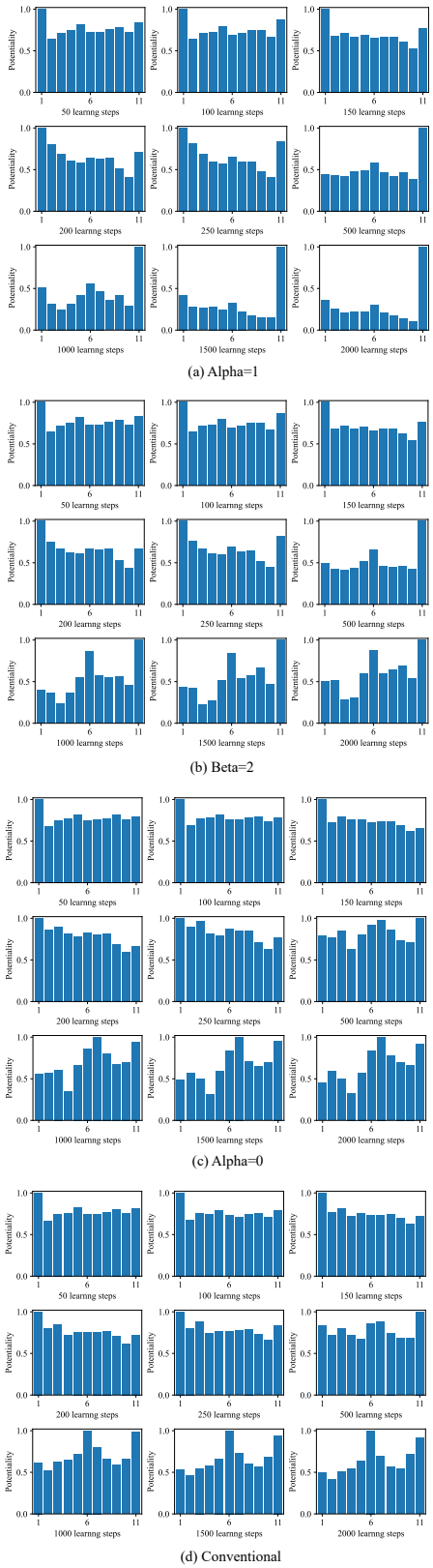


Figure 8. Layer potentialities for all layers, when the ratio potentiality was used (a), the diachronic potentiality was used (b), the distributed potentiality was used (c), and the conventional method was used (d).

suggests that, at first, local potentiality control attempted to detect the prototype network. By contrast, divergence in the middle and correlation coefficients did not clearly show this tendency. Generalization (right) increased rapidly at first and then decreased slightly toward the end. When the diachronic parameter was set to two with the best generalization in Figure 9(b), ratio potentiality continued to increase as the number of learning steps grew. The estimated prototype was forced to remain closer to the supposed prototype throughout learning. The divergence in the middle also decreased as learning progressed, though it could not be normalized and the tendency was not as clear. The right figure shows that generalization performance increased gradually, with a kind of phase transition occurring in the middle of learning. This indicates that improved generalization performance resulted from structural changes in weights achieved by controlling both local and distributed potentiality. When only the distributed potentiality was applied, as shown in Figure 9(c), ratio potentiality increased only at the beginning of learning, while divergence did not clearly reveal this tendency. Consequently, generalization (right) was lower across the entire learning process. When the conventional method was used, as in Figure 9(d), the ratio potentiality was initially high, but divergence decreased at first and then gradually increased later. Generalization (right), however, remained relatively low.

The results show that productive simplification consistently attempted to detect the prototype network throughout learning, whereas the other methods attempted detection only at the beginning. This persistent attempt by productive simplification seems to produce a kind of phase transition that improves generalization.

F. Interpreting Compressed Weights

The results show that under productive simplification, the weights were close to those of the supposed prototype network, which were obtained by computing the correlation coefficients between inputs and outputs independently. The best generalization performance was obtained by preserving the weights of the supposed prototype network. This means that improved generalization could be achieved through simplification.

Figure 10 shows the weights of compressed networks obtained by four methods. When only the local potentiality was used, as in Figure 10(a), the weights were initially close to those of the prototype, but gradually changed, with only the weight from input No. 3 remaining strong. When generalization was highest and the diachronic parameter was set to two in Figure 10(b), the weights in the initial stage of learning remained almost unchanged until the final stage. Productive simplification attempted to detect the supposed prototype as consistently as possible throughout learning. When only the distributed potentiality was used, as shown in Figure 10(c), the prototype was detected in the early stage, but later the weight from input No. 3 became dominant, with several other weights also relatively large. When the conventional method was used, as in Figure 10(d), the weight from input No. 3 was

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

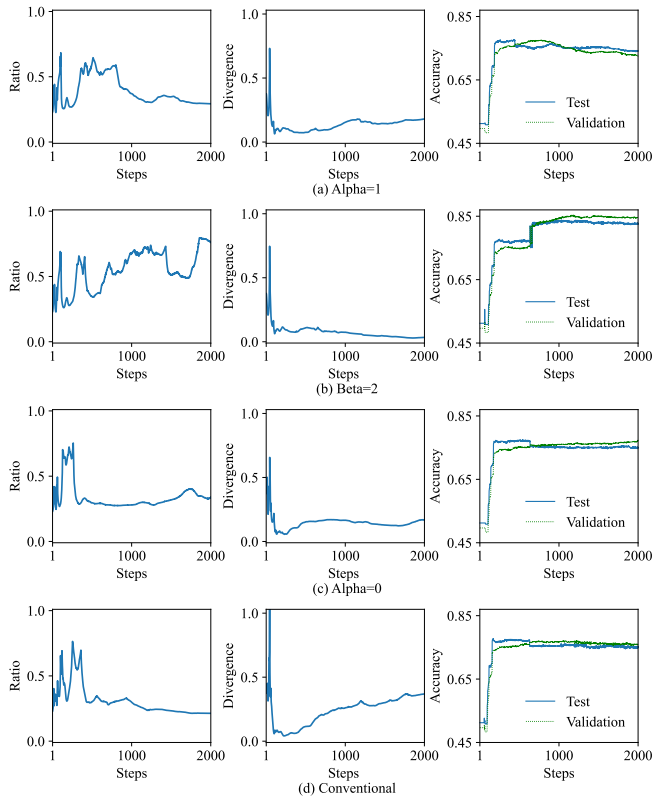


Figure 9. Ratio potentiality (left), divergence (middle), and generalization accuracy (right), when the local potentiality was used (a), the diachronic potentiality was used (b), the distributed potentiality was used (c), and the conventional method was used (d).

the largest, and all other weights were relatively weaker at the end.

The results show that the final connection weights obtained through productive simplification were very similar to those of the prototype network. This indicates that prototype detection or simplification was necessary even to increase generalization.

G. Numerical Summary

These numerical results reconfirmed that generalization was closely related to simplification. This suggests that behind seemingly complicated and nonlinear multilayer neural networks, there exists a simple prototype network that enables a surface network to improve generalization.

Table I summarizes the results, with the upper and lower tables representing the outcomes based on maximum ratio potentiality and generalization, respectively. The upper table shows the results when the ratio potentiality was the largest, where the final compressed networks were closest to the supposed prototype. For example, when the diachronic parameter was two, the ratio was the largest (0.796) with the largest number of learning steps (1857), the smallest simplification potentiality (0.126), and the highest generalization accuracy (0.827). Only the local potentiality was not so small (0.413), adding some complexity to the connection weights. This indicates that the best generalization was not obtained by simply decreasing the local potentiality, but by balancing local and distributed

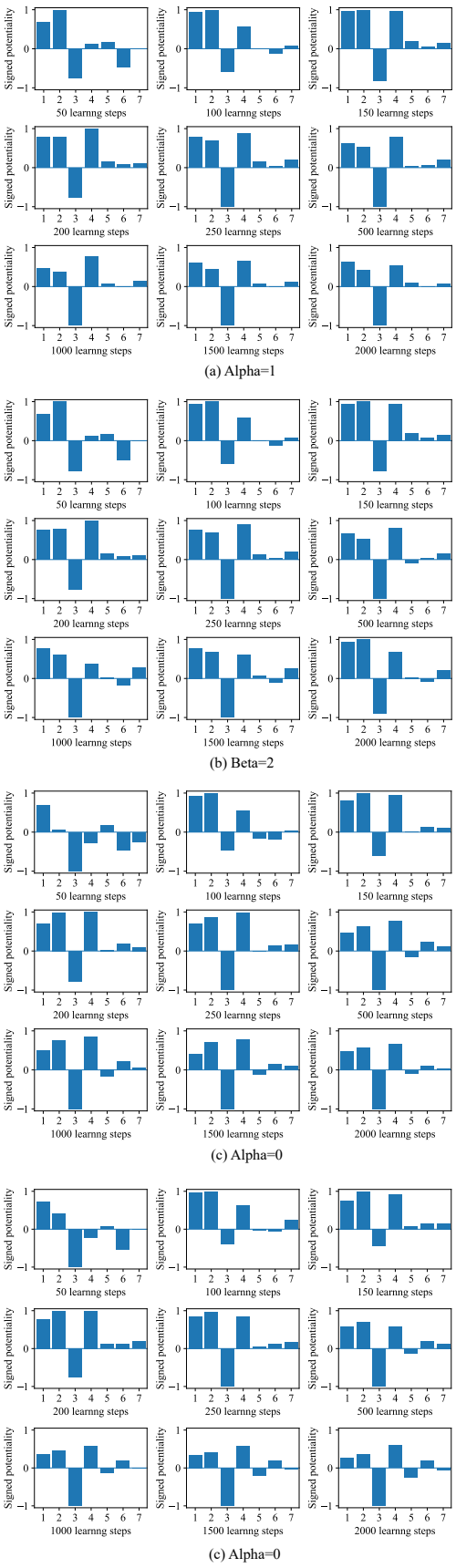


Figure 10. Weights of compressed networks, when the local potentiality was used (a), the diachronic potentiality was used (b), the distributed potentiality was used (c), and the conventional method was used (d).

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

potentiality. Even in this case, the total simplification was strictly maintained.

The lower table shows the results based on generalization performance. When the diachronic parameter was two, the generalization performance was the highest (0.838), and the ratio potentiality was also the largest (0.698) with the greatest number of learning steps (1032). The simplification potentiality was the second smallest (0.126), and the local potentiality was the second largest (0.301). Even when generalization performance was the criterion for evaluation, the ratio potentiality remained the largest and the total simplification potentiality the second smallest. The best generalization was obtained by controlling both simplification and local potentiality.

The results show that generalization could be improved by making networks as simple as possible, while allowing some complexity in the connection weights. This added complexity did not significantly increase the total simplification potentiality. These findings indicate that simplification can ultimately be controlled in correspondence with the necessary complexity.

TABLE I. SUMMARY OF RESULTS, BASED ON MAXIMUM RATIO POTENTIALITY (UPPER) AND MAXIMUM GENERALIZATION ACCURACY (LOWER).

Method	Ratio	Step	Simp	Local	Testing
Local	0.684	103	0.129	0.330	0.509
β_{dch}=2	**0.796**	**1857**	**0.126**	0.413	**0.827**
Distributed	0.754	260	0.176	**0.245**	0.768
Conventional	0.764	251	0.204	0.401	0.768

Method	Ratio	Step	Simp	Local	Testing
Local	0.277	254	**0.118**	**0.200**	0.776
β_{dch}=2	**0.698**	**1032**	0.126	0.301	**0.838**
Distributed	0.296	498	0.172	0.217	0.774
Conventional	0.341	166	0.195	0.404	0.777

V. CONCLUSION

The present paper aimed to show that there exists a strong simplification force inside all types of neural networks as a model of human cognition. In fact, there are different forms of simplification. For example, the number of components may be reduced in terms of local simplification, while the number of components may also be increased with reduced strength to capture the roles or features of each component. These different types of simplification can be unified productively to create a more powerful simplification force, enabling neural networks to handle a wider variety of new inputs. This productive simplification is naturally based on the simplification principle, but it can eventually produce some complexity in the connection weights. This simplified complexity generated by productive simplification is assumed to prevail in any cognitive system, including neural networks.

The method was applied to an artificial data set with linear and nonlinear inputs. The results show that two types of simplification could be unified productively to produce a more flexible simplification force. This productive unification could be realized through a phase transition, improving generalization performance. This means that productive simplification can drastically and structurally change network configurations for better generalization.

For future work, we should mention four points. First, we dealt only with the linear type of prototype network obtained by our method. Since our data set also contained nonlinear relations between inputs and outputs, we need to consider a nonlinear prototype network. Second, we need to extend our experiments to real-world data sets to examine how effectively our method can reveal the simplification force in neural networks. Third, the present method essentially attempted to simplify the well-known mutual information principle in neural networks, in which two contradictory operations of entropy maximization and minimization coexist. We therefore need to examine more precisely how our method can be used to realize the effect of mutual information maximization [21]. Fourth, we need to compare our simplification model with actual findings in neuroscience, where simplicity and complexity have often been discussed with some confusion. It is possible to extend our method to clarify the problem of simplicity in neuroscience [30], where simplification should play a more significant role.

REFERENCES

[1] R. Kamimura, "Neutralized synchronic and diachronic potentiality for interpreting multi-layered neural networks," in *Proceedings of the Seventeenth International Conference on Advanced Cognitive Technologies and Applications*, IARIA, 2025, pp. 17–25.

[2] T. Kohonen, "The self-organization map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[3] P. Schneider, M. Biehl, and B. Hammer, "Adaptive relevance matrices in learning vector quantization," *Neural Computation*, vol. 21, no. 12, pp. 3532–3561, 2009.

[4] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "Self-adaptive multiprototype-based competitive learning approach: A k-means-type algorithm for imbalanced data clustering," *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1598–1612, 2019.

[5] M. Zidan *et al.*, "Quantum classification algorithm based on competitive learning neural network and entanglement measure," *Applied Sciences*, vol. 9, no. 7, p. 1277, 2019.

[6] T. Shinozaki, "Biologically motivated learning method for deep neural networks using hierarchical competitive learning," *Neural Networks*, vol. 144, pp. 271–278, 2021.

[7] M. G. Mahdy, A. R. Abas, and T. M. Mahmoud, "Multi-phase adaptive competitive learning neural network for clustering big datasets," in *The International Conference on Artificial Intelligence and Computer Vision*, Springer, 2021, pp. 731–741.

[8] M. Biehl, B. Hammer, and T. Villmann, "Prototype-based models in machine learning," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 7, no. 2, pp. 92–111, 2016.

[9] F. Pourpanah *et al.*, "A review of generalized zero-shot learning methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[10] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *Advances in Neural Information Processing Systems*, vol. 31, pp. 7775–7784, 2018.

[11] C. Yang, A. Rangarajan, and S. Ranka, "Global model interpretation via recursive partitioning," in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, IEEE, 2018, pp. 1563–1570.

[12] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[13] G. Alain, "Understanding intermediate layers using linear classifier probes," *arXiv preprint arXiv:1610.01644*, 2016.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 1135–1144.

[15] D. Garreau and U. Luxburg, "Explaining the explainer: A first theoretical analysis of lime," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 1287–1296.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Stat*, vol. 1050, p. 9, 2015.

[17] P. Luo, Z. Zhu, Z. Liu, X. Wang, and X. Tang, "Face model compression by distilling knowledge from neurons," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[18] S. Hooker, A. Courville, G. Clark, Y. Dauphin, and A. Frome, "What do compressed deep neural networks forget?" *arXiv preprint arXiv:1911.05248*, 2019.

[19] R. Mishra, H. P. Gupta, and T. Dutta, "A survey on deep neural network compression: Challenges, overview, and solutions," *arXiv preprint arXiv:2010.03954*, 2020.

[20] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.

[21] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[22] ——, "Perceptual neural organization: Some approaches based on network models and information theory," *Annual Review of Neuroscience*, vol. 13, no. 1, pp. 257–281, 1990.

[23] ——, "Local synaptic learning rules suffice to maximize mutual information in a linear network," *Neural Computation*, vol. 4, no. 5, pp. 691–702, 1992.

[24] ——, "Improved local learning rule for information maximization and related applications," *Neural Networks*, vol. 18, no. 3, pp. 261–265, 2005.

[25] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, IEEE, 2015, pp. 1–5.

[26] A. M. Saxe *et al.*, "On the information bottleneck theory of deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124 020, 2019.

[27] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, 2020.

[28] R. K. Mahabadi, Y. Belinkov, and J. Henderson, "Variational information bottleneck for effective low-resource fine-tuning," *arXiv preprint arXiv:2106.05469*, 2021.

[29] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[30] E. Genç *et al.*, "Diffusion markers of dendritic density and arborization in gray matter predict differences in intelligence," *Nature Communications*, vol. 9, no. 1, p. 1905, 2018.

[31] M. M. V. Hulle, "The formation of topographic maps that maximize the average mutual information of the output responses to noiseless input signals," *Neural Computation*, vol. 9, no. 3, pp. 595–606, 1997.

[32] B. C. Geiger, "On information plane analyses of neural network classifiers—a review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7039–7051, 2021.

[33] C. Buciluă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2006, pp. 535–541.

[34] J. O. Neill, "An overview of neural network compression," *arXiv preprint arXiv:2006.03669*, 2020.

[35] K. Shimizu, *Multivariate Analysis (in Japanese)*. Nikkan Kogyo Shinbun, 2009.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Leveraging Graph-Centric Case-Based Reasoning for Enhancing Monotonicity & AI Coherency

Steve Chan
VTIRL, VT/DE-CAIR
Orlando, USA
stevec@de-cair.tech

***Abstract*—The Artificial Intelligence (AI) coherence (and concomitant AI hallucination) issue, which centers upon the issue of validity, among others, remains an ongoing challenge, and current mitigation methods have had limited efficacy for certain Large-Language Model (LLM)-based systems. This paper reviews the aforementioned points and presents some experimental conjoining of certain similarity measures along with AI-centric heuristic updating/generating in an endeavor to operationalize pathways that: (1) tend towards the monotonic arena, and (2) are, correspondingly, less likely to spawn towards the Non-deterministic Polynomial-time Hardness (NP-hard) non-continuous, non-polynomial, non-monotonic side.**

Keywords—Artificial Intelligence; Conversational AI Agents; AI Hallucinations; AI Coherence; Graph-Based Reasoning; Case-Based Reasoning; Analogical Reasoning; Isomorphism.

I. Introduction

This paper builds on [1], which covered Decision Quality (DQ) and certain Extrapolated Decision Quality (DQ) Thematics (EDQTs); DQ is a key underpinning element of Robust Dialogue Management (RDM), which is a non-trivial matter for Conversational AI Agents (CAs). For their conversational dialogues, CAs have the challenge of Sequential Decision-Making (SDM), which impacts maintaining logical flow, consistency, validity, and overall coherency during the course of not only a single conversation, but also, potentially, multi-turn conversations. This then segues into the matter of monotonic constraints, which are often looked to for the enforcement of logical flow and coherence. However, Real-World Scenarios (RWS) often tend to involve the non-monotonic side, wherein conclusions can be withdrawn/reversed/altered. While conventional reasoning involves deductive, inductive, and abductive forms, in contemporary times, Inductive Reasoning (IndR) seems to be among the more prevalently used, and there are Kahneman & Tversky underpinnings associated with this. When time is not of the essence, such as in a System 2 sense ("slow, deliberate, logical"), Deductive Reasoning (DedR) can be utilized for more complicated matters; however, when time is of the essence, such as in a System 1 sense (i.e., "fast, automatic, intuitive"), IndR might be more optimal. Within the category of IndR, Analogical Reasoning (AnaR) is often used, and within this, Case-Based Reasoning (CBR) seems to be among the more highly exercised. This constitutes an opportunity since AnaR is construed to be associated with the non-monotonic realm while CBR is cautiously monotonic; after all, the literature indicates that LLM-based CAs are likely to be better at example-based learning (e.g., CBR) rather than adhering to explicit guardrail or system prompt instructions. In turn, CBR can also be graph-based, thereby affording the opportunity to leverage Graph-Based Reasoning (GBR) and move from cautiously monotonic to monotonic. In this way, it is possible (for certain cases) to avail of the benefits of AnaR, via GBR/CBR, while staying more on the monotonic side.

This paper has been substantially re-written with new content (as contrasted to the conference paper). This paper is intended as a design paper, rather than as a complete empirical contribution, and focuses on reviewing the Artificial Intelligence (AI) coherence (and concomitant AI hallucination) issue, which centers upon the issue of validity, and delineates some of the current mitigation methods, which have had limited efficacy for particular Large-Language Model (LLM)-based systems. The paper presents some experimental conjoining of certain similarity measures along with AI-centric heuristic updating/generating in an endeavor to operationalize pathways that: (1) tend towards the monotonic arena, and (2) are, correspondingly, less likely to spawn towards the Nondeterministic Polynomial-time Hardness (NP-hard) noncontinuous, non-polynomial, non-monotonic side. Section I provided an overview regarding the challenges and complexities of RDM, which is a core requisite capability for CA. The remainder of the paper is organized as follows. Section II provides pertinent background information pertaining to the challenges of AI coherence (and hallucinations). Section III presents some theoretical foundations, experimental building blocks, and experimentation. Section IV provides a brief discussion and lists some of the limitations of the paper. Section V summarizes with concluding remarks, and proposed future work closes the paper.

II. Background

There are a variety of AI LLM leaderboards, but, among others, some of the more highly recognized LLMs include: OpenAI's Generative Pre-trained Transformer (GPT)-5, Alibaba's Tongyi Qianwen (Qwen)-3-235B-A22B, xAI's Grok-3, Meta's Llama 4, Deepseek's V3.1, Anthropic's Claude 3.7, and Google's Gemini 2.5 Pro [2]. These LLMs are leveraged for a myriad of applications, such as CA, which strives to undertake natural human conversation through various modes (e.g., text, voice). Some of the more popular CAs include Apple's Siri, ChatGPT, Gemini, and ElevenLabs AI 2.0. Yet, these CAs are

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

beset by the technical issue of what is often referred to as "AI hallucinations," which OpenAI deems to be "plausible but false statements" and IBM deems to be "nonsensical or altogether inaccurate" outputs [3][4].

A. The Dilemma and Challenge

Against this backdrop of AI hallucinations, the referenced CA market is aggressively growing. The market research firm Grand View Research asserts that the CA "market size was estimated at USD 11,576.4 million in 2024 and is projected to reach USD 41,393.2 million by 2030, growing at a [Compound Annual Growth Rate] CAGR of 23.7% from 2025 to 2030" [5]. The firm Markets and Markets seems to be aligned with this assertion and notes that CA "is projected to be USD 49.80 billion by 2031, growing from 17.05 million in 2025" [6]. The firm Fortune Business Insights is even more aggressive in its assertion: [the CA] "market size is expected to grow from $12.24 billion in 2024 to $61.69 billion in 2032" [7]. In support of the described CA consumer market, Verified Market Research asserts that the CA "platform software market was valued at $234.82 million in 2024 and is projected to reach $589.76 million by 2031" [8]. As a summarization, the various market research firms cite tremendous growth for the CA ecosystem. Therein lies the dilemma; while CA consumer demand is accelerating (along with the underpinning platform infrastructure ecosystem), CA has been beset with technical issues, such as that of the referenced AI hallucinations.

Kate Irwin of *PC Magazine* reports on an interview that the *Washington Post* had with Apple CEO Tim Cook, wherein the takeaway was that "Apple's AI tools may not always be accurate" and may be subject to "AI hallucinations" [9]; William Gallagher of *AppleInsider* adds to this by noting that AI hallucinations may occur much more frequently than expected/desired [10]. A plethora of researchers have published studies regarding the frequency of AI hallucinations for CA, and in the case of Athaluri (as an exemplar), it is reported that ChatGPT's AI hallucinations resulted in (out of the 178 references cited): 28 fictitious references and 41 erroneous Digital Object Identifiers (DOIs) (i.e., a 39% hallucination rate) [11]. *Tech Crunch* has reported that "according to OpenAI's internal tests, o3 and o4-mini, which are so-called reasoning models, hallucinate *more often* than the company's previous reasoning models — o1, o1-mini, and o3-mini — as well as OpenAI's traditional, 'non-reasoning' models, such as GPT-4o" [12]. *Mashable* follows up on this by noting that while "o3's hallucination rate is 33 percent,"…"o4-mini's hallucination rate" (for a "more advanced version") is much higher at "48 percent" [13]. OpenAI, in its technical report, entitled "Open AI o3 and o4-mini System Card," notes that "more research is needed" as to why there are increasingly "more inaccurate/hallucinated claims" as the versions advance and as reasoning models are scaled up [14]. Shifting to Google, it self-notes that Gemini "can sometimes hallucinate," and Imad Khan at *CNET* notes that "Gemini can be slow, prone to hallucinate and links to incorrect pieces of information" [15][16]. Hugging Face further comments that the actual hallucination rate for the "latest Gemini model" may be higher than Google's reported rates [17]. In an endeavor to address this matter, during the AI Action Summit of February 2025, Google DeepMind and Giskard announced the Potential Harm Assessment & Risk Evaluation (PHARE) LLM benchmark, and the associated paper reports that "leading LLM systems consistently struggle with [AI] hallucination, exhibiting high variability across different contexts" [18][19][20]. Alan Weissberger remarks on PHARE, via the *IEEE Communications Society*, and notes that "some models" have "hallucination rates exceeding 30% in specialized fields [21]. Along this vein, reportage based upon the "RealHarm Dataset" of "problematic interactions" with CA indicates that "misinformation and fabrication represent approximately one-third of documented incidents, confirming that [AI] hallucination remains the primary challenge in production LLM systems despite significant research attention" [22][23]. Perhaps, OpenAI put it the best on 5 September 2025: "Even as language models become more capable, one challenge remains stubbornly hard to fully solve: [AI] hallucinations" [24].

B. Conventional Mitigation Approaches

Human oversight/validation is an axiomatic mitigation approach for addressing AI hallucinations. However, putting aside certain "high-stakes"/mission-critical applications (e.g., law, healthcare), the Human-in-the-Loop (HITL) approach does not necessarily scale well or at speed, as noted by Holzinger and others [25]. Apart from HITL, there are a variety of more automated approaches, among others, for addressing AI hallucinations: (1) Guardrails/System Prompts (GSP), (2) Fine-Tuned Models (FTM) for specialized domains, and (3) Real-time Retrieval Augmented Generation (RAG). These mitigation approaches are discussed in 1) through 3).

1) Guardrails/System Prompts (GSPs)

Andrew Cunningham of *Ars Technica* reminds us of existing system prompts, which — even as acknowledged by Apple itself — may not rise to the desired levels of mitigation against AI hallucinations; some of these system prompts are a constituent part of various metadata.json files within the "System/Library/AssetsV2/com_apple_MobileAsset_UAF_FM_GenerativeModels/purpose_auto folder on Macs running the macOS Sequoia 15.1 beta that have also opted into the Apple Intelligence beta" [26]. Upon examination, mitigating system prompts include: "'do not hallucinate,' 'do not make up factual information,' etc." [26]. Moving from system prompts to guardrails, there are OpenAI guiding principles (in their "Cookbook"), such as: "provide very descriptive metrics to evaluate whether a response is accurate," "ensure consistency across key terminology," "evaluate each sentence independently and then entire response as a whole," etc. [27]; on the Amazon front, guardrails include using a "contextual grounding check policy" to "detect and filter AI hallucinations in model responses that are not grounded in enterprise data" [28]. In short, while system prompts can serve as shaping operations to help ensure better alignment with the envisioned accuracy/precision and contextual relevance, guardrails can serve as a Quality Assurance/Quality Control (QA/QC) mechanism for the output of the system prompts.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

2) Fine-Tuned Models(FTMs) for Specialized Domains

FTMs can indeed constitute a potential mitigation strategy for reducing AI hallucinations by leveraging/applying a pre-trained model to a particular "high-quality dataset" in the hopes of anchoring AI outputs via domain-specific knowledge (e.g., declarative, procedural, conditional, etc.) [29]. By way of background, declarative knowledge (e.g., recitals of fact, concepts) is "knowing what," procedural knowledge (e.g., step-by-step skills that are utilized/actions that are instinctively performed via "implicit memory" or "muscle memory") is "knowing how," and conditional knowledge is "knowing when and why" to strategically apply declarative and procedural knowledge [30][31][32]. Declarative knowledge can be encapsulated in both Known and Unknown forms (e.g., facts that exist but are not yet learned, facts that cannot be immediately recalled and are temporarily "unknown," "difficult-to-verbalize" recitals of fact/notions that are deemed to be implicit, etc.); likewise, procedural knowledge can be in both Known and Unknown (i.e., implicit) forms. In other words, when handling the "Unknown Unknowns" (UU) and "Known Unknowns" (KU) of the KU, UU, Unknown Knowns (UK), and "Known Knowns" (KK) epistemological (pertaining to the theory of knowledge) model leveraged by Shaker and Moore-Clingenpeel as well as others (and popularized by Donald Rumsfeld), such as shown in Table I below, there exists a dramatic distinction (as pertains to the rate of AI hallucinations) between the Knowns and the Unknowns.

TABLE I. EPISTEMOLOGICAL CONSTRUCTS [33][34]

Known Knowns (KK) "Things we are aware of and understand"	*Known Unknowns (KU)* "Things we are aware of and do not understand"
Unknown Knowns (UK) "Things we are not aware of, but understand"	*Unknown Unknowns (UU)* "Things we are not aware of and do not understand"

While FTMs can better mitigate against AI hallucinations on Known (e.g., KK) information, when contending with Unknown information (e.g., KU, UK, UU), they can, potentially, aggravate the AI hallucinations paradigm. To further accentuate this, Gekhman's Sampling-based Categorization of Knowledge (**Sli**C**K**) Known/Unknown model, which is a variant of Table I, is shown in Table II, wherein the last column, *Resultant*, depicts how often greedy decoding predicts the correct answer. As part of the setup for Section III Experimentation herein, Gekhman's study setup is adopted. Hence, "given a fine-tuning dataset D and a pre-trained LLM M," M_D denotes "a model obtained by fine-tuning M on D" [35]. Gekhman had adopted the perspective that "M knows that the answer to q is a[,] if it generates a when prompted to answer q," and Gekhman also defined $P_{correct}(q,a;M,T)$ "as an estimate of how likely is M to accurately generate the correct answer a to q, when prompted with random few-shot exemplars and using decoding temperature T " [35][36][37].

TABLE II. GEKHMAN'S SLICK MODEL [35]

Type	***Category***	***Definition***	***Resultant***
Known	"Highly Known" (HK)	"$P_{correct}(q,a;M,T=0)=1$"	"Always"
	"Maybe Known" (MK)	"$P_{correct}(q,a;M,T=0) \in (0,1)$"	"Sometimes"
	"Weakly Known" (WK)	"$P_{correct}(q,a;M,T=0)=0 \wedge P_{correct}(q,a;M,T>0)>0$"	"Never with T=0, but Sometimes with T>0"
Unknown	"Unknown"	"$P_{correct}(q,a;M,T\geq 0)$"	"Never"

According to Gekhman's ascertainment as to the negative impact of Unknown examples, the finding is that a "higher Unknown ratio is proportional to performance degradation" [35]. Restated, "LLMs struggle to learn new factual information through unsupervised fine-tuning" and this can result in unanticipated/undesirable outcomes (e.g., a higher AI hallucination rate) [38]. Furthermore, as noted by Sun, "a common challenge when fine-tuning LLMs for domain-specific applications is the potential degradation of the model's generalization capabilities" [39]. However, it has been reported that complementary methods, such as Retrieval-Augmented Generation (RAG), may assist with improving performance reliability (e.g., accuracy).

3) Retrieval-Augmented Generation (RAG)

To obviate against simply relying upon internal static training data, RAG can connect to external dynamic sources of data. This helps to contend with potentially out-of-date training data and can, hopefully, enhance the overall contextual awareness for the purpose of lessening the probability of AI hallucinations. Yet, RAG is also plagued by issues, such as the retrieval of specious and/or irrelevant information as well as the occasional erroneous fusing/leveraging of external information.

C. An Alternative Approach: Moving from AI Hallucinations to AI Coherence

It should be noted that AI hallucinations and AI coherence are distinct and disparate notions (albeit interrelated). While AI hallucinations are a generated response consisting of inaccurate or specious information, AI coherence refers to the "logical consistency" of the output response. The conventional approaches toward the AI hallucination challenge (described in Section IIB) have had limited efficacy (as spotlighted in the latter part of Section IIA). Accordingly, this paper addresses matters upstream of AI hallucinations; hence, AI coherence is treated. The treatment of AI coherency is significant, as a collapse/degradation of coherence can be a harbinger of AI hallucinations (yet, not all AI hallucinations involve incoherence, as an AI hallucination may have logical flow, but still be false). Maintaining coherence (via, theoretically, stringent monotonic constraints) amidst RWS is a complicated matter, "as the CA might discern connections (particularly those that are non-monotonic) within the ever-evolving dataset, and non-monotonic facets may materialize as incoming information re-contextualizes and/or contradicts matters;" moreover, "enforcing a strict monotonic paradigm can segue to an unnatural rigidity and/or incorrect/irrelevant responses by the CA." Prior research has shown that enhanced insight into the CA

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

behavior at Monotonic/Non-monotonic Transition Zones (MNTZ) can, potentially, be quite meaningful for enhancing coherency and consistency (with the concomitant validity). Yet, the treatment of MNTZ is often not part of contemporary CA architectures. Accordingly, Section III will unpack this further.

III. Experimentation

A. Theoretical Foundations

1) Reasoning Mechanisms (RMs)

For CA, the primary RMs, by validity ranking, are likely to be DedR, IndR, and then AbdR. Grote-Garcia states that DedR "is the process of using general premises to draw specific conclusions" (i.e., a "top-down paradigm") [40]. In contrast, for IndR, conclusions are derived by progressing from the specific to the general (i.e., a "bottom-up paradigm"); the University of Illinois Springfield puts it nicely: "inductive reasoning is the ability to combine pieces of information that may seem unrelated to form general rules or relationships" [41]. According to Minnameier, despite the recognized creative aspects (as Holyoak reminds us) of AnaR, it "is widely conceived of" as IndR (which is considered to be non-creative) [42][43]. In terms of validity ranking, AnaR ("the ability to perceive and use relational similarity between two situations or events") is subordinately situated below IndR, but it sits above AbdR (which may start with "puzzling observations" and segue to "inferring the most likely explanations") [44][45]. AnaR and AbdR may also leverage secondary RMs, such as CBR, while the former may also utilize GBR. Yan describes CBR as being "based on the cognitive assumption that similar problems have similar solutions" while Taylor & Francis describes CBR as being "a problem-solving approach that involves using past successful solutions to similar problems to solve new problems" [46][47]. Kolodneer notes that CBR is an AnaR method, and accordingly, CBR is subordinately situated below AnaR, as shown in Table III [48].

TABLE III. Types of primary reasoning mechanisms (RMs) [49]

Information Available		***RM***	***Resultant***
Examples: " • Facts • Accepted Truths • Rules • Scientific Laws	• Mathematical Theorems • Established Principles • Logical Connections"	DedR	"Always true, if the premises and arguments are valid."
"Starts with the same set as Deductive Reasoning (if available) and involves a probabilistic approach."		IndR	"Likely to be true but could be false despite the observations being accurate."
		AnaR	
		CBR	
"Starts with the same set as Deductive Reasoning (if available), but also involves hypotheses, assessments, and best-fit approximations."		AbdR	"Sometimes true, as it is a plausible best guess approximation."

Meanwhile, GBR is a technique that buttresses reasoning capabilities by characterizing problems as graphs and then proceeding to resolve the problems via the ascertaining of connections and patterns. The various primary and secondary RMs discussed can also be sorted by the involved Reasoning Processes (RPs) (i.e., monotonic, non-monotonic).

2) Reasoning Processes (RPs)

As noted in prior work, RPs include Monotonic Reasoning (MR) and Non-Monotonic Reasoning (NMR). MR responses "remain consistent throughout time despite whatever new information might arrive" while NMR responses "allow for modification and/or retraction of prior assertions." As the thematics and priorities within the conversational dialogue change and as new information may alter the context and/or contradict prior information, the addressing of RDM and "MR and NMR becomes critical for maintaining consistency and interconnectedness, particularly for multi-turn conversations;" "if the constituent elements of a multi-turn conversation are indeed logically related, then the overall dialogue should be relatively free of contradictions," and this constitutes a paradigm of "conversational coherence." To validate conversational coherence, there are a variety of evaluation tools that can be leveraged. For example, the "ConversationCoherence Evaluator" is "a tool designed to check the coherence of conversations by an AI," as "it evaluates whether each response in a conversation logically follows from the previous messages, ensuring that the AI maintains context and relevance throughout the interaction" [50][51]. However, "'conversational coherence' is, for CA, quite difficult to maintain because the information supply changes temporally, and at some points, it may be sparse/incomplete and/or ambiguous/uncertain." Depending upon "what" and "when" the information is made available, a particular RM(s) may be more apropos. Table IV, which was reviewed in [49], provides a sampling of RMs as well as their associated RP (i.e., MR/NMR) categorization, which is illuminated using a Red-Orange-Yellow-Green (ROYG) color coding schema, wherein MR is indicated by green, NMR by red, weak MR by orange, and cautious MR by yellow [49][52][53].

TABLE IV. RM-centric and RP-centric sorting [49]

RM	***MR/NMR Categorization***	
DedR	*MR*	
GBR	*MR*	*NMR*
CBR	*Cautious[ly] MR*	*NMR*
IndR	*Weak[ly] MR*	*NMR*
AnaR	*NMR*	
AbdR	*NMR*	

Table IV is quite interesting as the often used AnaR is categorized as NMR. As can be seen, CBR is cautious[ly] MR while GBR is MR. Hence, if there is an opportunity to leverage GBR or GBR/CGR, then there is potential to stay more within the MR realm (i.e., higher coherence).

3) CA, IndR, AnaR, CBR, and GBR

As a follow-on, Chen points out that "one of the intriguing abilities of LLMs is reasoning" [54]. As an extension to this, "one of the intriguing abilities of" CA is also reasoning, which consists of the primary RMs of Table III: DedR, IndR (with its subordinate AnaR and CBR), and AbdR. While Chen asserts that IndR is central to LLMs[/CAs], in Chen's study, it was

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

found that LLM performance was somewhat sub-optimal for IndR [54]. Similarly, Luo reports that, when using the LogiGLUE benchmark (which is comprised of 24 datasets consisting of DedR, IndR, and AbdR), "the findings indicate that LLMs excel most in" AbdR, "followed by" DedR, "while they are least effective" for IndR [55]. Along this vein, Cheng points out that "while the" DedR "capabilities of LLMs, (i.e. their capacity to follow instructions in reasoning tasks), have received considerable attention, their abilities in true" IndR "remain largely unexplored" [56]. Cheng also claims that "LLMs demonstrate remarkable" IndR "capabilities through SolverLearner," but Eliot cautions that "depending upon how the generative AI was devised by an AI maker, such as the nature of the underlying foundation model, the capacity to undertake" IndR varies greatly [56][57]. Overall, "the findings suggest that LLMs might be better at learning by example and discovering patterns in data than at following explicit instructions," thereby hinting at the value-added proposition of CBR (and AnaR) [58]. Leake suggests that CBR provides a "basis for learning from few examples" (few-shot), and Qin asserts that AnaR can be used "to address unfamiliar challenges by transferring strategies from relevant past experiences"/examples [59][60]; as a reminder, CBR is a specialized form of AnaR [49][61].

CBR involves a similarity-based comparison, which might include various key features. An ensuing more robust comparison might entail finding cases/examples that are structurally identical and involve an isomorphic-based comparison; this can be achieved, such as by way of an Isomorphic Paradigm (IsoP) Comparator Similarity Measure (CSM) (ICSM), and prior work in this regard has shown that graph-based CBR (i.e., a GBR/CBR amalgam) can be quite suitable for IsoP. As pertains to the CSM, in some cases, such as for unordered sets, the ordering of the edges (and their weights) may not necessarily be relevant, as only the nodes and their values need to be compared; for example, when providing the temperature (T), relative humidity (RH), and wind speed (WSPD) (along with their associated values) for a particular point in time, the sequencing of T, RH, and/or WSPD is not of any particular import. In other cases, such as for ordered sets, the edges and the sequencing of the nodes is of significance, such as the [node] stops that are to be made along a delivery route. Some CSM considerations (e.g., Pre-IsoP, IsoP) from the prior work of [49] are presented in Table V, sample Partially Ordered Sets (POSETs) are shown in Table VI, and exemplar Isomorphic Variants (IVs) are shown in Table VII.

TABLE V. CSM (PRE-ISOP AND ISOP) CONSIDERATIONS [49]

Considerations	*Definition*
Unordered Set (UnS)	"A set of disparate constituents, wherein the order of the constituents is not relevant. By way of example," {T, RH, WSPD} equates to {WSPD, T, RH} and "{1, 2, 3, 4, 5} equates to {5, 3, 1, 4, 2}."
Equal Sets (EqS)	"A pair of sets S and S' is equal if and only if (iff) each constituent of S is also a constituent of S'; moreover, the order of the constituents is not relevant. By way of example, if S = [1, 2, 3, 8, 9, 10] and S' = {9, 3, 1, 2, 10, 8}, then S=S'."
Equivalent Sets (EquivS)	"A pair of sets S and S' is considered equivalent if the number of constituents in S and S' is the same (i.e., same cardinality). By way of example, if S = {1, 3, 5, 7, 9} and S' = {2, 4, 6, 8, 10}, then S and S' are considered to be equivalent."
Ordered Set (OrS)	"A set of disparate constituents, wherein the order of the constituents is relevant, and the constituents can be ordered and compared via operators, such as by <. By way of example, an ordered set might be {1, 2, 3, 5, 6, 8, 9, 10}, whereas an unordered set might be {6, 5, 1, 2, 3, 10, 9, 8}."
Partially Ordered Set (POSET)	"A set of disparate constituents, wherein the constituents might or might not be able to be ordered and compared, since operators such as <= can yield different variations. By way of example, Calcworkshop (https://calcworkshop.com/relations/partial-order/) provides some examples, which we extrapolate upon in the way of {a < b < c < d <= e <= f }, {a <= b <= c < d < e <= f < g}, and {a < b < c <= d <= e < f}, which are shown diagrammatically in Table VI."
Unordered Sets with Isomorphism (UnS-Iso)	"For a set of disparate constituents, wherein the constituents are unordered, if there is a one-to-one relationship (i.e., bijection), then the unordered sets are likely isomorphic. By way of example, if S={1, 2, 3, 4, 5}, S'={a, b, c, d, e}, and 1<->a, 2<->b, 3<->c, 4<->d, and 5<->e (wherein each constituent in S relates to a unique constituent in S'), then S and S' are considered to be isomorphic."
POSETs with Isomorphism (POSET-Iso)	"For a set of disparate constituents, wherein the constituents are considered to be within a POSET, if there is a bijection, then the POSETs are likely isomorphic. By way of example, if S={S_1, S_2, S_3, S_4, S_5}, S'={S'_2, S'_4, S'_5, S'_3, S'_1}, and S_1<->S'_2, S_2<->S'_5, S_3<->S'_1, S_4<->S'_4, and S_5<->S'_3, (wherein each constituent in S relates to a unique constituent in S'), then S and S' are considered to be isomorphic. To demonstrate this, online tools are available, such as https://graphonline.top/en/?graph=xPLjwOkrglDRgYe, among others."
Isomorphism Variants (IV)	"Extrapolating upon the POSETs with Isomorphism (POSET-Iso), there are also permutations that are actually isomorphism variants (e.g., automorphism, which is a particular type of isomorphism that has a symmetrical structure), which Lemons nicely depicts and for which examples are shown in Table VII as IV#1 through #3" [62].

TABLE VI. ISOMORPHIC PARTIALLY-ORDERED SETS (POSETS) [49]

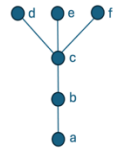
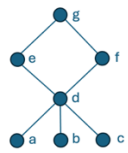
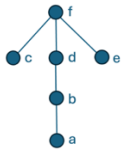
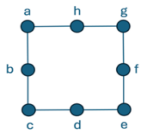
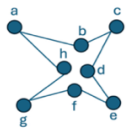
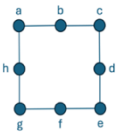
POSET #1	*POSET #2*	*POSET #3*
		
Maximal: d, e, f	Maximal: g	Maximal: f
Greatest: none	Greatest: g	Greatest: f
Minimal: a	Minimal: a, b, c	Minimal: a
Least: a	Least: none	Least: a

TABLE VII. ISOMORPHIC VARIANTS (IVS) [49]

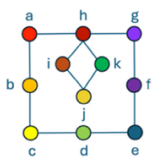
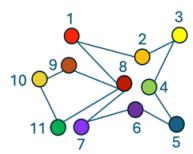
IV #1	*IV #2*	*IV #3 (automorphism)*
		

With regards to GBR, if there exists a one-to-one correspondence between the vertices of S and S', then S and S'

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

are isomorphic. This is shown in Table VIII (which was previously presented in [49]), and this can be affirmed via a variety of tools, such as the one available at https://graphonline.top/en/?graph=xPLjwOkrglDRgYeS. Also, rather than the graphs themselves, adjacency matrices can also be utilized to determine isomorphism. This can be affirmed via a variety of tools, such as the one available at https://graphonline.top/en/create_graph_by_matrix. Given this versatility, the leveraging of graph-based CBR can be quite advantageous; in fact, when GBR and CBR are conjoined (i.e., GBR/CBR), it is possible to, potentially, move towards the green MR side (as contrasted to the yellow cautious[ly] MR side) of Table IV. Li, Xu, and many others seem to be a proponent of this approach [63][64]. By moving from the yellow to the green, it is likely contributory towards reducing the propensity for spawning towards the NP-hard, non-continuous, non-polynomial, and NMR side.

TABLE VIII. EXEMPLAR ISOMORPHISM BETWEEN S AND S' [49]

S and S' Isomorphism	*Graph S*	*Graph S'*
f(a) = 1 *f(b)* =2 *f(c)* =3 *f(d)* =4 *f(e)* =5 *f(f)* =6 *f(g)* =7 *f(h)* =8 *f(i)* =9		

It would also, likely, enhance AnaR via the GBR/CBR amalgam. This is important, as analogies are prevalent in conversation; metaphors are also prevalent. As a quick primer, simile "is a comparison of two disparate entities, via words, such as 'like' or 'as'," metaphor "is a direct comparison and asserts that two disparate entities are the same, via words, such as 'is,' 'was,' etc. (wherein the words 'like' or 'as' are not utilized), analogy "creates a comparison of how a seemingly disparate entity is akin to, relates to, or is similar to another disparate entity for the purpose of explaining/demonstrating," and allegory "embodies a more complex/symbolic comparison and leverages a narrative to convey an abstract notion/concept." In a study by Casarett, it was found that "the use of metaphors and analogies may enhance physicians' ability to communicate," as metaphors appeared in 64% of the conversations, while analogies were used in 31% of the conversation, and on average, doctors used 1.6 metaphors and 0.6 analogies per conversation [65]. According to Kanthan, similes, metaphors, and analogies "bridge the Known to the Unknown" [66]; hence, in accordance with Table II, the utilization of metaphors and analogies (which facilitates the movement from the Unknown to the Known) segues to a paradigm, wherein the resultant validity is likely to be higher (with regards to Table II, the validity for HK is "always," MK is "sometimes," and WK is "sometimes" when T>0). Moreover, the handling can be accomplished by CBR or GBR/CBR rather than simply AnaR; the key distinction here is that GBR/CBR reside in the green/yellow of MR and cautious[ly] MR, respectively, while AnaR resides in the red of NMR.

B. Experimental Building Blocks

Pragmatically, GBR/CBR can leverage various similarity metrics for the determination of similar cases/examples. An alphabetized sampling of measures is shown in Table IX, and these were then sorted by their linear/non-linear paradigms and their categorization with regards to monotonicity/non-monotonicity, such as shown in Table X; this builds upon the work from [67].

TABLE IX. VARIOUS MEASURES FOR MONOTONIC/NON-MONOTONIC AND LINEAR/NON-LINEAR PARADIGMS [67]

Measure	*Descriptor*
Distance Correlation Coefficient (dCor) [67]	"dCor is 'better at revealing complex…relationships…compared with other correlation metrics' by 'integrating both linear and non-linear dependence'" [71].
Hoeffding's D Correlation Coefficient (D) [67][68]	"D can reflect a certain degree of concordance and discordance."
Information Coefficient of Correlation (ICC) [69]	"ICC can provide a gauge of alignment between the posited and actual value."
Kendall's Tau Correlation Coefficient (tau) [67][70]	"Tau can illuminate correlations of significance when the distributions of the sample set and population are not necessarily known."
Maximal Correlation (MC) [69]	"MC pertains to transformations of the data, which are considered to maximize the correlation."
Maximum Information Coefficient (MIC) [69]	"MIC encompasses both linear and nonlinear correlations between the 'variable pairs'."
Mutual Information (MI) [69]	"MI is a paradigm, wherein one of the variables conveys a quantifiable amount of information about the other."
Pearson's [Product]-Moment Correlation Coefficient (PPMCC) [67]	"PPMCC measures the relationship strength and direction between the 'variable pairs'."
Percentage Bend Correlation Coefficient (PBCC) [67]	"PBCC refers to a paradigm, wherein a specified percentage of marginal observations deviating from the median are weighted downward" [72].
Spearman's Rho Correlation Coefficient (rho) [67][70]	"Rho scrutinizes the dependence between two random variables" [73].

TABLE X. EXEMPLAR USAGE OF VARIOUS MEASURES [67]

		Monotonic	*Non-monotonic*
Linear		D [67] rho [67] tau [67][70] PPMCC [67][69][70] PBCC [67] dCor [67]	N/A[2]
Non-linear		PPMCC[1][67][69]70] rho [67] tau [67][70] PBCC [67] dCor [67] D [67]	MC [69] dCor [67] D [67] PPMCC[3] [67][69] Rho[3] [67][69]
	Curvilinear	rho [67] PBCC [67] dCor [67] PPMCC[1][67] tau [67]	dCor [67] D [67] PPMCC[3] [69] Rho[3] [67][69]

[1] "Heuvel notes the efficacy of PPMCC with 'families of bivariate distribution functions with non-linear monotonic associations'" [70].

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[2] "Technically, non-monotonic cannot be linear; however, as noted by Nicolaou, linear dynamics may experience transient segueing 'toward non-monotonic dynamics'" [74].
[3] "Of note, given symmetry, it does not 'find non-monotonic dependence'" [69].

The experimentation conducted yielded results that seemed to align with the findings of Mirtagioglu [67]. For example, the following seem to hold: (1) "in cases where there is no relationship (i.e., 0) between the variables" (e.g., non-functional relationship, wherein "there is no function of one variable that interacts with the other and vice versa"), dCor, D, tau, PPMCC, PBCC, rho, as well as MC "have given very satisfactory results," (2) "very low values (i.e., close to 0)" of rho, tau, PPMCC, and PBCC is emblematic of a "random relationship between the variables," and (3) "very low values (i.e., close to 0)" of tau, PPMCC, and PBCC, and rho when conjoined with "very high values (close to 1)" of dCor is emblematic of a non-monotonic relationship between/among variables, such as shown in Table XI [67][70].

TABLE XI. EXEMPLAR FINDINGS FROM MEASURES & POSITS [67]

Close to 0	*Close to 1*	*Close to -1*	*Relationship*
dCor, D, tau, PPMCC, PBCC, rho	N/A	N/A	None
rho, tau, PPMCC, PBCC	N/A	N/A	Random
N/A	rho, PPMCC	N/A	Strong Positive Monotonic
N/A	N/A	rho, PPMCC	Strong Negative Monotonic
tau, PPMCC, PBCC, rho,	dCor	N/A	Non-monotonic

The results also somewhat align with the findings of Rainio, and Heuvel. However, the rankings and sortings, such as offered by Mirtagioglu (M), Rainio (R), and Heuvel (H) somewhat differ, as shown in Table XII below. In terms of computational complexity, D is at O(n log n) while dCor is at $O(n^2)$ [75][76]; hence D is faster than dCor, and these are the validation measures for NMR. Moreover, D is less sensitive to outliers and can be well suited for ties. Next, rho is at O(n log n) while tau can be between O(n log n) and $O(n^2)$ [77][78][79]; hence, rho can be faster than tau for handling non-linear MR (however, tau tends to be less sensitive to outliers and can be well suited for ties). PBCC is considered to be at O(n log n), but it is subject to the involved sorting algorithms. MIC is at $n^{2.4}$[80]. Finally, PPMCC can be considered to be between O(n) and O(n log n). Hence, PPMCC can be faster than rho (in some cases) (however, rho tends to be less sensitive to outliers and has higher efficacy than PPMCC). Using a ROYG color-coding schema, the complexities are shown in color within Table XII, which stems from the work of [67].

TABLE XII. POSITED RANKING/SORTINGS BY M, R, AND H [67]

		M [68]	***R*** [70]	***H*** [71]
Monotonic	***Linear***	rho PBCC PPMCC dCor tau	PPMCC[1] rho[2] tau[2]	PPMCC MIC
	Non-linear	rho PBCC dCor tau D	rho tau PPMCC	PPMCC MIC
Non-monotonic	***Non-linear (e.g., curvilinear)***	dCor D	N/A	PPMCC[3] rho[3] MIC

[1] "more oriented for 'linear association'" [71].
[2] "more oriented for 'monotonic association'" [71].
[3] "however, this is N/A when the non-monotonic dependence is symmetric" [71].

Other Similarity Measures (SMs) include the Heterogeneous Euclidean-Overlap Metric (HEOM) and Value Difference Metric (VDM), which can accommodate situations with mixed numerical and categorical data [82][83]; for example, dimensions constitute numerical data, but colors are categorical data. Measures, such as these as well as others, will be considered in future work. For the experiments herein, the Table XIII complexity considerations for the SMs of Tables IX-XII were utilized. Then, those SMs that were the most computationally tractable (and more on the monotonic side), for both the initial foray and validation shown in Figure 1 (which builds upon the work of [67]), were arranged into Lower-Level Heuristic (LLH) amalgams, as shown in Table XIV.

TABLE XIII. COMPLEXITY CONSIDERATIONS FOR SMS

O(n)	*O(n)* <=SM<= *O(n log n)*	*O(n log n)*	*O(n log n)* <=SM<= *O(n2)**	*O(n2)**	*O(n2.4)**
	PPMCC	D rho PBCC	tau	dCor	MIC

*for sufficiently large values of n

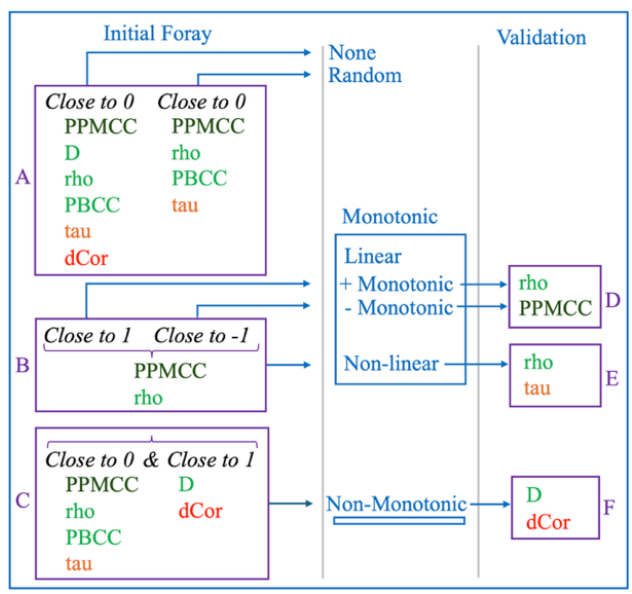


Fig. 1. Construct Utilized for Experimental Runs [67]

TABLE XIV. LLH AMALGAMS FOR EXPERIMENTAL RUNS

LLH A	***SM(s) A***	***LLH B***	***SM(s) B***	***LLH D***	***SM(s) D***
A1	PPMCC	B1	PPMCC	D1	PPMCC
A2	rho	B2	rho	D2	rho
A3	PBCC	***LLH C***	***SM(s) C***	***LLH E***	***SM(s) E***
A4	D-rho	C1	PPMCC-D	E1	rho

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A5	D-PBCC	C2	rho-D	***LLH F***	***SM(s) F***
		C3	PBCC-D	F1	D

These LLHs are conjoined alongside the Hyper-Heuristics (HH) and Metaheuristics (MH) experimented with in prior work. This is shown in Figure 2, which is an extrapolation of Bouazza's work [84].

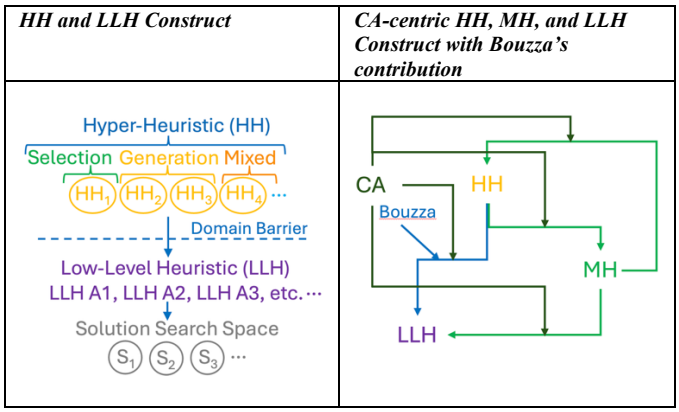


Fig. 2. Construct Utilized for Experimental Runs [85]

Experimentation was conducted on open LLM models, such as Mistral AI's Mixtral 8x7B (Apache 2.0 license) (LLM1), Berkeley's Neural Engineering Systems Technology's (NEST) Starling-LM-7B-Alpha (Apache 2.0 license with the additional condition that the model is not used to compete with OpenAI) (LLM2), and Mintplex Lab's AnythingLLM (MIT license) (LLM3, which used both LLM1 and LLM2). The relative performance from the experimental runs of Table IV and Figure 2 are shown in Figure 3.

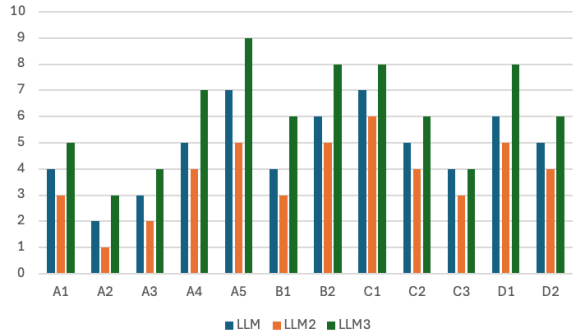


Fig. 3. LLH Performance Results from Experimental Runs

E1 and F1 were excluded, as they were solitary LLHs. LLM1 seemed to exhibit superior results to LLM2, and interestingly, LLM3 performed better than LLM1. LLM1's Sparse Mixture of Experts (SMoE) might have accounted for its performance, and LLM3's seemingly successful amalgamation of LLM1 and LLM2 warrants further investigation. The performance of LLM A4 and A5 was not surprising given D, and for this set of experimental runs, PBCC seemed to perform better than rho. In the case of the LLH C set, PPMCC seemed to perform better than rho and PBCC; along this vein, the LLH D set emulates this trend (however, LLH B set does not). Future work will involve further experimentation in this regard.

Then, as an extrapolation of Table IV, Table XIII is derived. Under an Uncompressed Decision Cycle (UDC) paradigm (more akin to System 2), DedR can be utilized. Under a Compressed Decision Cycle (CDC) paradigm (more akin to System 1), AbdR or IndR can be used; using a ROYG color coding schema, green has the fastest relative performance while red has the slowest relative performance. In essence, as DedR is more analytical, it tends to be slower; as IndR endeavors to establish a pattern, it tends to be slower than AbdR, which can be faster in putting forth an "inference to the best explanation." As IndR tends to be prevalent for CA, AnaR, CBR, and GBR are compared. As can be seen in Table XIII, given comparable performances, it seems prudent to opt for GBR given that it has a heightened probability of staying within the green MR zone.

TABLE XV. RMs AND RPS UNDER UDC AND CDC

UDC		***CDC***		
DedR	*MR*	IndR	*W MR*	*NMR*
		AnaR	*NMR*	
		CBR	*C MR*	*NMR*
		GBR	*MR*	*NMR*
		AbdR	*NMR*	

IV. Discussion

This paper advances a design framework, wherein which GBR-facilitated CBR and order-theoretic formalisms (e.g., posets, isomorphisms, monotonicity) are leveraged to facilitate the assessment of coherence in LLM outputs. Among other aims, one research Line of Effort (LOE) explores whether model outputs can be mapped to structured representations whose relations are partially ordered (i.e., in a poset fashion), whereby desirable reasoning behavior can be delineated as monotone movement within that order and quantified with designated rank and/or dependence statistics.

The paper depicts how certain cases, exemplars, and their relations might be represented as nodes and order relations, and it leverages the use of correlation and/or rank concordance measures (e.g., Pearson, Spearman, Kendall, Hoeffding) as prospective diagnostics of monotonic progress. As a limitation (and intended as future work), this particular paper does not explicitly specify the operational pipeline; this will be presented in a follow-on piece of work that is already in-progress. That work will also clearly map the limitations of the various measures used. For example, Pearson correlation assumes linear dependence and is sensitive to scale, Spearman and Kendall are able to capture monotone associations, but do not suffice in affirming causal monotonicity in reasoning, and Hoeffding's measure can have high efficacy in some cases but is also data-hungry and can be unstable with small samples. As can be gleaned, this work-in-progress will contain a lengthy survey section with known limitations. Future work will be further discussed in Section V.

V. Concluding Remarks

RWS tend to be more in the NMR realm. Consequently, enforcing MR constraints to maintain coherence for CA is

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

challenging, and achieving RDM is non-trivial. This AI coherence issue tends to devolve to incidents of AI hallucinations, and despite various contemporary mitigation approaches (GSP, FTM, RAG, etc.), the problem seems to be, as reported in the literature, worsening as CA versions advance. In essence, as opined by a number of researchers in the arena, the treatment of the AI hallucination/AI coherence issue has been sub-optimal for various LLM-based CA systems; AI hallucinations, AI coherence, and validity seem to be linked to the state of the information-at-hand. For the case of the Known (e.g., KK, particularly HK, as contrasted to MK or WK), the validity seems to remain higher (relatively speaking). For the case of the Unknown (e.g., KU, UK, UU), the validity declines. Of course, the decline in validity is accompanied by a dramatic performance degradation with regards to AI coherence (and hallucinations). Generally speaking, CA conversations occurring in real-time, from the CA vantage point, tend to reside somewhere between CDC and UDC. Hence, the involved primary RM would likely tend more towards IndR (rather than AbdR or DedR, respectfully). Along this vein of IndR (and its subordinate oft used AnaR), GBR/CBR seems to be optimal. Moreover, in the context of CA, Zheng and others have long asserted: to "build emotional bond with users," more "advanced linguistic features" should be incorporated [86]. Within the literature, historical studies (e.g., Glucksberg, Kaall, Roberts) have shown "that figurative language…are key to interesting and engaging conversations" [86]. In addition, Hofstadter has argued that analogy is the "core of cognition," and Holyoak seems to affirm [87][88]. Hofstadter further states, "without concepts there can be no thought, and without analogies there can be no concepts" [89]. At its core, the conveyance of related/similar concepts/notions is central for RDM. Accordingly, the secondary RMs of CBR and GBR become significant; after all, LLMs are "optimally suited at learning by example," and the issue then becomes how to best operationalize matters. While there is of course, interest in assessing the involved computational complexity, the goal of staying as much in the MR realm (as opposed to NMR), amidst computational practicality, rises to the forefront (and the contextualization of the MNTZ becomes crucial). The experimentation within demonstrates that there are some heuristic amalgams (involving certain similarity measures) that seem to be apropos for this goal. Future work will involve more quantitative and qualitative experimentation in addition to that proposed in Section IIIB.

REFERENCES

[1] S. Chan, "Treatment of the Multi-Attribute Decision-Making Rank Reversal Problem for Real-World Systems," *The Seventeenth International Conference on Future Computational Technologies and Applications (Future Computing)*, Apr. 2025, pp. 1-10.

[2] "Tongyi Qianwen (Qwen)," *Alibaba Cloud*, [Online]. Accessed: Sep. 5, 2025. Available: https://www.alibabacloud.com/en/solutions/generative-ai/qwen?_p_lc=1.

[3] A. Kalai, O. Nachum, S. Vempala, and E. Zhang, "Why Language Models Hallucinate," *Arxiv.org*, Sep. 4, 2025 [Online]. Accessed: Sep. 5, 2025. Available: https://arxiv.org/abs/2509.04664.

[4] "What are AI hallucinations," *IBM*, Sep. 1, 2023. [Online]. Accessed: Sep. 1, 2025. Available: https://www.ibm.com/think/topics/ai-hallucinations.

[5] "Global Conversational AI Market Size & Outlook, 2024-2030," *Grand View Horizon*, [Online]. Accessed: Sep. 1, 2025, Available: https://www.grandviewresearch.com/horizon/outlook/conversational-ai-market-size/global.

[6] "Conversational AI Market worth $49.80 billion by 2031," *Markets and Markets*, [Online]. Accessed: Sep. 1, 2025. Available: https://www.marketsandmarkets.com/PressReleases/conversational-ai.asp.

[7] "Conversational AI Market Size," *Fortune Business Insights*, Aug. 25, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://www.fortunebusinessinsights.com/conversational-ai-market-109850.

[8] "Conversational AI Platform Software Market Size and Forecast," *Verified Market Research*, [Online]. Accessed: Sep. 1, 2025. Available: https://www.verifiedmarketresearch.com/product/conversational-ai-platform-software-market/.

[9] K. Irwin, "Tim Cook: Apple Intelligence Isn't Immune to Hallucinations," *PC Magazine*, Jun. 11, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://www.pcmag.com/news/tim-cook-apple-intelligence-isnt-immune-to-hallucinations-wwdc.

[10] W. Gallagher, "Siri Chatbot prototype nears ChatGPT quality, but hallucinates more than Apple wants," *AppleInsider*, Jun. 1, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://appleinsider.com/articles/25/06/01/siri-chatbot-prototype-nears-chatgpt-quality-but-hallucinates-more-than-apple-wants.

[11] S. Athaluri et al., "Exploring the Boundaries of Reality: Investigating the Phenomenon of Artifical Intelligence Hallucination in Scientific Writing Through ChatGPT References," *Cureus*, vol. 15, pp. 1-5.

[12] M. Zeff, "OpenAI's new reasoning AI models hallucinate more," *TechCrunch*, Apr. 18, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://techcrunch.com/2025/04/18/openais-new-reasoning-ai-models-hallucinate-more/.

[13] C. Mauran, "OpenAI's o3 and o4-mini hallucinate way higher than previous models," *Mashable*, Apr. 19, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://mashable.com/article/openai-o3-o4-mini-hallucinate-higher-previous-models.

[14] "OpenAI o3 and o4-mini System Card," *OpenAI,* Apr. 16, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

[15] K. Anish, "Getting the Most from Gemini: Understanding its Knowledge and Creativity," *Gemini Apps Help*, Nov. 26, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://support.google.com/gemini/community-guide/309961349/getting-the-most-from-gemini-understanding-its-knowledge-and-creativity?hl=en.

[16] I. Khan, "Google Gemini Chatbot Review: Hallucination Station," *Cnet*, Apr. 2, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://www.cnet.com/tech/services-and-software/google-gemini-chatbot-review-hallucination-station/.

[17] A. Zilber, "Google's AI is 'hallucinating,' spreading dangerous info — including a suggestion to add glue to pizza sauce," *New York Post*, Jun. 6, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://nypost.com/2025/06/06/business/googles-ai-overviews-are-hallucinating-by-spreading-false-info/.

[18] M. Dora, "Giskard announces Phare, a new open & multi-lingual LLM Benchmark," *Giskard*, [Online]. Accessed: Sep. 1., 2025, Available: https://www.giskard.ai/knowledge/giskard-announces-phare-a-new-llm-evaluation-benchmark#:~:text=This%20new%20LLM%20benchmark%2C%20called,samples%20will%20be%20open%2Dsource.

[19] "Phare: A Safety Probe for Large Language Models," *Hugging Face*, May 16, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://huggingface.co/papers/2505.11365.

[20] P. Jeune, B. Malezieux, W. Xiao, and M. Dora, "Phare: A Safety Probe for Large Language Models," *Arxiv.org*, May 26, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/abs/2505.11365.

[21] A. Weissberger, "Sources: AI is Getting Smarter, but Hallucinations Are Getting Worse," *IEEE ComSoc Technol.*, May 10, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://techblog.comsoc.org/2025/05/10/nyt-ai-is-getting-smarter-but-hallucinations-are-getting-worse/.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[22] P. Jeune, J. Liu, L. Rossi, and M. Dora, "RealHarm: A Collection of Real-World Language Model Application Failures," *RealHarm Dataset*, [Online]. Accessed: Sep 4, 2025. Available: https://realharm.giskard.ai/.

[23] P. Jeune, J. Liu, L. Rossi, and M. Dora, "RealHarm: A Collection of Real-World Language Model Application Failures," *Proc. of the The First Workshop on LLM Security (LLMSEC)*, Aug. 2025, pp. 87-100.

[24] "Why language models hallucinate," *OpenAI*, Sep. 5, 2025. [Online]. Accessed: Sep. 5, 2025. Available: https://openai.com/index/why-language-models-hallucinate/.

[25] A. Holzinger, K. Zatloukal, and H. Muller, "Is human oversight to AI systems still possible?" *New Biotechnology*, vol. 85, pp. 59-62, Mar. 2025.

[26] A. Cunningham, ""Do not hallucinate": Testers find prompts meant to keep Apple Intelligence on the rails," *ArsTechnica*, Aug. 6, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://arstechnica.com/gadgets/2024/08/do-not-hallucinate-testers-find-prompts-meant-to-keep-apple-intelligence-on-the-rails/.

[27] R. Ziv, "Developing Hallucination Guardrails," *OpenAI Cookbook*, May 29, 2024. [Online] Accessed: Sep. 1, 2025, Available: https://cookbook.openai.com/examples/developing_hallucination_guardrails.

[28] A. Gutpa, "Use Guardrails to prevent hallucinations in generative AI applications," *AWS*, Jul. 11, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://builder.aws.com/content/2i12ntqFx3xAaDLfvrjH7278sEW/use-guardrails-to-prevent-hallucinations-in-generative-ai-applications.

[29] "Domain specific knowledge and skills," *ScienceDirect*, [Online]. Accessed: Sep. 1, 2025. Available: https://www.sciencedirect.com/topics/psychology/domain-knowledge.

[30] A. Huaulme, "Surgical declarative knowledge learning: concept and acceptability study," *Comput. Assisted Surgery*, vol. 27, pp. 74-83, 2022.

[31] L. Fisher, B. Halima, and K. Yerian, "Procedural and Declarative Knowledge," *Learning How to Learn Languages*, 2024 [Online]. Accessed: Sep. 5, 2025. Available: https://opentext.uoregon.edu/languagelearningedition1/chapter/procedural-and-declarative-knowledge/.

[32] I. Zsigmond, "Role of Conditional Knowledge in Conscious Reading: The Integrating Model of Metacognition," *Proc. of the 16th European Conf. on Reading and 1st Ibero-American Forum on Literacies*, Jan. 2009, pp. 1-8.

[33] M. Shaker and M. Moore-Clingenpeel, "The known knowns, known unknowns, and unknown unknowns of surveys and sleep," *Ann. Allergy Asthma Immunol*., vol. 129, pp. 669-670, Dec. 2022.

[34] S. McGregor, "Learning with Donald Rumsfeld – flexible learning: the relevance and resonance of multiprofessional learning in primary care," *Br. J. Gen Pract*., vol. 54, pp. 722-723, Sep. 2004.

[35] Z. Gekhman et al., "Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?" *Proc. of the 2024 Conf. on Empirical Methods in Natural Lang. Process*., Nov. 2024, pp. 7765-7784.

[36] S. Kadavath et al., "Language Models (Mostly) Know What They Know," *Arxiv.org*, Nov. 21, 2022. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/pdf/2207.05221.

[37] P. Mankul, A. Liusie, M. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," *Proc. of the 2023 Conf. on Empirical Methods in Natural Lang.Process.*, Dec. 2023, pp. 9004-9017.

[38] O. Elisha, "Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs," *Arxiv.org*, Jan. 2024. [Online]. Accessed: Sep. 14, 2025. Available: https://arxiv.org/pdf/2312.05934.

[39] J. Sun et al., "Dial-insight: Fine-tuning Large Language Models with High-Quality Domain-Specific Data Preventing Capability Collapse," *Arxiv.org*, Mar. 14, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/abs/2403.09167.

[40] S. Grote-Garcia, "Deductive Reasoning," in *Encyclopedia of Child Behavior and Development*. Boston, MA. Springer, 2011, pp. 477-478.

[41] "Inductive Reasoning," *University of Illinois Springfield, ION Professional eLearning Programs*, [Online]. Accessed: Sep. 1, 2025. Available: https://www.uis.edu/ion/resources/oiai/inductive-reasoning.

[42] G. Minnameier, "Abduction, Induction, and Analogy," in *Model-Based Reasoning in Science and Technology. Studies in Computational Intelligence*, vol. 314, pp. 107-119, 2010.

[43] K. Holyoak, N. Ichien, and H. Lu, "Analogy and the Generation of Ideas," *Creativity Research Journal*, vol. 36, pp. 532-543, Jul. 2023.

[44] D. Gentner and L. Smith, "Analogical Reasoning," in *Encyclopedia of Human Behavior*. Oxford, UK. Elsevier, 2012, pp. 130-136.

[45] A. Sandoval-Hernandez and D. Rutkowski, "Embracing complexity: abductive reasoning as a versatile tool for analyzing international large-scale assessments," in *Educ. Assessment Eval. And Accountability*, vol. 37, pp. 255-271, Dec. 2024.

[46] A. Yan and Z. Cheng, "A Review of the Development and Future Challenges of Case-Based Reasoning," *Appl. Sci.*, vol. 14, pp. 1-22, Aug. 2024.

[47] "Case-based reasoning," *Taylor & Francis*, [Online]. Accessed: Sep. 1, 2025. Available: https://taylorandfrancis.com/knowledge/Engineering_and_technology/Artificial_intelligence/Case-based_reasoning/.

[48] J. Kolodner, "Improving Human Decision Making through Case-Based Decision Aiding," *AI Magazine*, vol. 12, pp. 52-68, 1991.

[49] S. Chan, "Interstitial b-SHAP-Owen Amalgam for the Enhancement of Artificial Intelligence System-Centric Sequential Decision-Making," *International Journal on Advances in Intelligent System*s, v18, in press.

[50] "How to Evaluate AI Chats Using Conversation Coherence Evaluator," *Athina AI*, Apr. 17, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://blog.athina.ai/how-to-evaluate-ai-chats-using-conversation-coherence-evaluator.

[51] "Athina-ai/athina-evals," *Github*, [Online]. Accessed: Sep. 1, 2025. Available: https://github.com/athina-ai/athina-evals/blob/main/examples/conversation_coherence.ipynb.

[52] K. Jantke, "Monotonic and non-monotonic inductive inference of functions and patterns," *Lecture Notes in Computer Science*, vol. 543, pp. 161-177, Jan. 2005.

[53] G. Paulino-Passos and F. Toni, "Monotonicity and Noise-Tolerance in Case-Based Reasoning with Abstract Argumentation," *Proc. of the 18th Int. Conf. on Princ. of Knowl. Representation and Reasoning*, Nov. 2021, pp. 508-518.

[54] B. Chen, R. Saetre, and Y. Miyao, "A Comprehensive Evaluation of Inductive Reasoning Capabilities and Problem Solving in Large Models," *Findings of the Assoc. for Comput. Linguistics*, pp. 323-339, Mar. 2024.

[55] M. Luo et al., "Towards LogiGLUE: A Brief Survey and A Benchmark for Analyzing Logical Reasoning Capabilities of Language Models," *Arxiv.org*, Mar. 31, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/html/2310.00836v3.

[56] K. Cheng, "Inductive or Deductive? Rethinking the Fundamental Reasoning Abilities of LLMs," *Arxiv.org*, Aug. 7, 2024. [Online]. Accessed: Sep. 5, 2025. Available: https://arxiv.org/abs/2408.00114.

[57] L. Eliot, "On Whether Generative AI And Large Language Models Are Better At Inductive Reasoning Or Deductive Reasoning And What This Foretells About The Future Of AI," *Forbes*, Aug. 11, 2024, [Online]. Accessed: Sep. 1, 2025. Available: https://www.forbes.com/sites/lanceeliot/2024/08/11/on-whether-generative-ai-and-large-language-models-are-better-at-inductive-reasoning-or-deductive-reasoning-and-what-this-foretells-about-the-future-of-ai/.

[58] B. Dickson, "LLMs excel at inductive reasoning but struggle with deductive tasks, new research shows," *Venture Beat*, Aug 15, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://venturebeat.com/ai/llms-excel-at-inductive-reasoning-but-struggle-with-deductive-tasks-new-research-shows.

[59] D. Leake and D. Crandall, "On Bringing Case-Based Reasoning Methodology to Deep Learning," *Case-Based Reasoning Methodology to Deep Learning. Lecture Notes in Computer Science*, vol. 12311, pp. 343-348, Oct. 2020.

[60] C. Qin et al., "Relevant or Random: Can LLMs Truly Perform Analogical Reasoning?" *Findings of the Assoc. for Comput. Linguistics*, Jul. 2025, pp. 23993-24010.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[61] M. Voskoglou and A. Salem, "Analogy-Based and Case-Based Reasoning: Two sides of the same coin," *Int. J. of Appl. Of Fuzzy Sets and Artif. Intell*., vol. 4, pp. 5-51, 2014.

[62] N. Lemons, B. Hu, W. Hlavacek, "Hierarchical graphs for rule-based modeling of biochemical systems," *BMC Bioinformatics*, vol. 12, pp. 1-13, Feb. 2011.

[63] J. Li, X. Luo, and G. Lu, "GS-CBR-KBQA: Graph-structured case-based reasoning for knowledge base question answering," *Expert Syst. with Appl.*, vol. 257, pp. 1-37, Dec. 2024.

[64] H. Xu, Y. Wei, Y. Cai, and B. Xing, Knowledge graph and CBR-based approach for automated analysis of bridge operational accidents: Case representation and retrieval," *PLOS One*, Nov. 2023, pp. 1-21.

[65] D. Casarett, "Can Metaphors and Analogies Improve Communication with Seriously Ill Patients," *J. Palliat Med*., vol. 13, pp. 1-6, Mar. 2010.

[66] R. Kanthan and S. Mills, "Using Metaphors, Analogies and Similes as Aids in Teaching Pathology to Medical Students," *Medical Sci. Educ*., vol. 16, pp. 102-116, Dec. 2017.

[67] S. Chan, "A Prospective Monotonic/Non-Monotonic Transition Zone Impediment for Concept Model-Centric Artificial Intelligence Systems," *The Second International Conference on AI-based Systems and Services (AISys)*, in press.

[68] H. Mirtagioglu and M. Mendes, "On Monotonic Relationships," *Biostatistics and Biometrics*, vol. 10, pp. 1-11, May 2022.

[69] A. Fujita, J. Sato, M. Demasi, "Comparing Pearson, Spearman, and Hoeffding's D measure for gene expression association analysis," *J. of Bioinformatics and Comput. Biology*, vol. 7, pp. 663-684, Sep. 2009.

[70] O. Rainio, "Different Coefficients for Studying Dependence," *The Indian J. of Stat*., vol. 84-B, pp. 895-914, Nov. 2022.

[71] E. Heuvel and Z. Zhan, "Myths About Linear and Monotonic Associations: Pearson's r, Spearman's ρ, and Kendall's τ," *The Amer. Stat.*, vol. 76, pp. 44-52, Nov. 2021.

[72] J. Hou, et al., "Distance correlation application to gene co-expression network analysis," *BMC Bioinformatics*, vol. 23, pp. 1-24, Feb 2022.

[73] C. Pernet, R. Wilcox, and G. Rousselet, "Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox," *Front. Psychol*., vol. 3, pp. 1-18, Jan. 2013.

[74] Y. Dodge, "Spearman Rank Correlation Coefficient," In: *The Concise Encyclopedia of Statist.* New York, NY: Springer, 2008, pp. 502-505.

[75] Z. Nicolaou, T. Nishikawa, S. Nicholson, J. Green, and A. Motter, "Non-normality and non-monotonic dynamics in complex reaction networks," *Phys. Rev. Res. 2*, pp. 1-15, Oct. 2020.

[76] C. Even-Zohar, "Independence: Fast Rank Tests," *Arxiv.org*, Oct. 2020. [Online]. Accessed: Sep. 5, 2025. Available: https://arxiv.org/pdf/2010.09712.

[77] J. Hou et al., "Distance correlation application to gene co-expression network analysis," *BMC Bioinformatics*, vol. 23, pp. 1-24, Feb. 2022.

[78] M. Stephanou and M. Varughese, "Sequential estimation of Spearman rank correlation using Hermite series estimators," *Arxiv.org*, Jul. 2021. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/pdf/2012.06287.

[79] M. Sepulveda, "Kendallknight: An R Package for Efficient Implementation of Kendall's Correlation Coefficient Computation," *Arxiv.org*, Dec. 8, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/pdf/2408.09618.

[80] D. Christensen, "Fast algorithms for the calculation of Kendall's τ," *Comput. Stat*., vol. 201, pp. 51-62, Mar. 2005.

[81] F. Shao and H. Liu, "The Theoretical and Experimental Analysis of the Maximal Information Coefficient Approximate Algorithm," *De Gruyter*, Mar. 2021, pp. 1-10.

[82] C. Li and H. Li, "Correlation weighted heterogeneous euclidean-overlap metric," *Int. J. of Comput*., vol. 33, pp. 341-346, Jul. 2015.

[83] L. Jiang and C. Li, "Two improved attribute weighting schemes for value difference metric, " *Knowledge and Inf. Syst*., vol. 60, pp. 949-970, 2019.

[84] W. Bouazza, "Machine Learning-Based Hyper-Heuristics: A Clear Insight," *Proc. of the 2024 7th Int. Conf. on Comput. Intell. And Intell. Syst. (CIIC)*, Feb. 2025, pp. 29-37.

[85] S. Chan, "AI-Centric Hyper-Heuristic and Self-Exclusion Mechanism for the Updating of a Heuristic," *IEEE World AI IoT Congress (AIIoT)*, May 2025, pp. 0536-0545.

[86] D. Zheng, R. Song, T. Hu, H. Fu, J. Zhou, "Love is as Complex as Math: Metaphor Generation System for Social Chatbox," *Chinese Lexical Semantics. Lecture Notes in Comput. Sci.*, vol. 11831, pp. 337-347, Jan. 2020.

[87] D. Hofstadter, "Epilogue: Analogy as the Core of Cognition," in *The Analogical Mind*. Cambridge, MA. MIT Press, 2001.

[88] K. Holyoak, "The Human Edge: Analogy and the Roots of Creative Intelligence," Cambridge, MA. MIT Press, pp. 39-41, 2025.

[89] D. Hofstadter and E. Sanger, "Surfaces and essences: Analogy as the fuel and fire of thinking," *Basic Books*, pp. 1-578, April 2013.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Interstitial b-SHAP-Owen Amalgam for the Enhancement of Artificial Intelligence System-Centric Sequential Decision-Making

Steve Chan
VTIRL, VT/DE-CAIR
Orlando, USA
Email: stevec@de-cair.tech

***Abstract*—The leveraging of an Interstitial b-SHAP-Owen Amalgam (IbSOA) is the focus of this paper; interestingly, the amalgam can potentially lend towards the enhancement of Artificial Intelligence (AI) System (AIS)-Centric multi-stage Sequential Decision-Making (DM) (SDM) amidst Ambiguity and Uncertainty (A/U) by leveraging certain previously peer-reviewed preset modules — Lower Ambiguity, Higher Uncertainty (LAHU) and Higher Ambiguity, Lower Uncertainty (HALU) Module (LHM) — and certain subordinate modules — LHM's Isomorphic Paradigm (IsoP) Comparator Similarity Measure Module (ICSM2) and LHM's Metaheuristic Algorithm Module (MAM) — as well as a new module in the form of a Conversational AI Robustness (CAIR) Accelerant (CAIR-A) Module (as a proxy case study example); CAIR-A addresses the challenge of Robust Dialogue Management (RDM) with the objective of sufficiently supporting a Conversational AI Agent (CAA) so as to be able to maintain consistency, coherency, and validity throughout an ongoing dialogue. To achieve this, the involved Reasoning Mechanisms (RMs) — Monotonic Reasoning (MR) and Non-Monotonic Reasoning (NMR) — need to be well counterpoised; the apropos harmonizing of MR/NMR can help ensure the RDM's plasticity for enhanced conversational coherence. In turn, the successful counterpoising of MR/NMR is predicated upon insights into the Monotonic/Non-Monotonic Transition Zones (MNTZs); after all, shifts from a monotonic to a non-monotonic paradigm and vice versa in the MNTZs can occur at a higher than anticipated rate. IbSOA-related insights into the myriads of interplays among local, glocal, and global facets can better contextualize the behavior within the MNTZs and better support the CAA's SDM amidst A/U. As previously noted, the CAA is simply a proxy application, and the implications of the referenced counterpoising have broader implications for AIS SDM.**

Keywords-artificial intelligence systems; machine learning; LAHU; HALU; LHM; IsoP; Monotonic Reasoning; Non-Monotonic Reasoning; MNTZ; Interstitial b-SHAP-Owen Amalgam (IbSOA); Sequential Decision-Making (SDM).

I. INTRODUCTION

Conversational AI Agents (CAA) endeavor to emulate natural human conversation through, among other modes, text and/or voice (CAA capability has also been extended to include holographic form, such as offered by Holoconnects, Ravatar, Proto, and others). Other companies focus on certain realism aspects, and offerings include those by Hume (https://www.hume.ai/?tab=evi), ElevenLabs (https://elevenlabs.io/conversational-ai), etc. Conventional CAA leverages an amalgam of constituent technologies, such as Automatic Speech Recognition (ASR), Natural Language Processing (NLP), Natural Language Understanding (NLU), Natural Language Generation (NLG), Large Language Models (LLMs), and Machine Learning (ML), among others, so as to engage in meaningful dialogue and provide robust "human-like responses." In contrast to prototypical chatbots, which are underpinned by "rule-based" prescribed scripts, CAA can engage in unscripted dialogues, learn/tune from each and every engagement, as well as leverage external sources/systems. In terms of realism, CAA should be able to perform in real-time without any awkward delays. This paper builds upon [1], which was submitted to Future Computing 2025 on March 1, 2025 (now published), as well as touches upon certain elements of [2], which was submitted to AI-based Systems and Services (AISyS) 2025 on June 10, 2025 (accepted; in press), and fundamentally, this paper, which was substantially completed in April 2025 (following the journal invitation), underscores the notion that CAA responses need to remain consistent, coherent, and valid. This is non-trivial to achieve, and a number of technical challenges exist within this ecosystem. Among other challenges, first, the CAA must contend with the SDM challenge; in essence, the series of decisions that the CAA makes will have a cumulative effect as to how robust the overall conversation (which may ensue across multiple interactions) will be. Second, the CAA needs to undertake SDM amidst varying degrees of A/U, whose definitions are provided in Table I below.

TABLE I. AMBIGUITY (A) VERSUS UNCERTAINTY (U)

DM Notions	*Definitions*
Ambiguity (A)	Ambiguity is prevalent when the involved circumstance(s) and/or the involved *information* can be construed in varying ways (including contradictory).
Uncertainty (U)	Uncertainty is prevalent when the outcome of the involved circumstance(s) is unknown/unpredictable, and there exists a lack of knowledge (wherein *information* pertains to recitals of fact and/or descriptors while *knowledge* pertains to the comprehension and operationalization of that *information*).

Interestingly, the issue of SDM extends far beyond the CAA case study, as will be delineated in Section IA.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A. The SDM Challenge

Autonomous Systems (AS), such as in the form of Autonomous Artificial Intelligence Systems (AIS), may include Autonomous Underwater Vehicles (AUVs), Unmanned Surface Vehicles (USVs), Unmanned Aerial Vehicles (UAVs), unmanned spacecraft, etc. A more commonly recognized manifestation might be in the form of self-driving cars. Some of the involved companies are quite well-known in the marketplace: Amazon's Zoox, General Motors' Cruise, Tesla, and Waymo, with the latter two, in particular, asserting their desire to achieve the Society of Automotive Engineers' (SAE) sixth level of driving-related automation – "5 (full automation)," which is contrasted to "0 (no automation)." This type of sixth level autonomous AI-centric vehicle, as well as certain "near-to-this-level" paradigms — as pertains to the other referenced drones — are necessarily reliant upon robust Sequential Decision-Making (DM) (SDM), which is often referred to as multi-stage DM since the SDM challenge itself resides in the fact that multiple DM points — over an elongated temporal span (i.e., a longer time horizon) — are interrelated. Each decision made at a DM point potentially constitutes a DM Inflection Point (DMIP) and impacts the ensuing prospective possible courses of action and their outcomes, such as described by Wanke, for which an abridged version is shown in Figure 1 below [3].

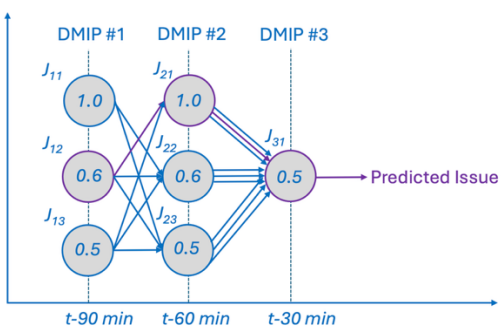


Figure 1. Various DM points comprising an exemplar SDM problem

Utilizing Wanke's presented case, 1.0 equates to "no resolution actions will be taken," 0.5 equates to "resolve the anticipated issue," 0.6 equates to a "partial resolution," J_{ij} equates to "resolution cost distribution at decision point i of option j," and the mean cost for the indicated path in purple of 0.6-1.0-0.5 is expressed in (1) [3].

$$E(J)=J_{12}+E(J_{21}|J_{12})+E(J_{31}|J_{21}|J_{12}) \qquad (1)$$

It should be clear from Figure 1 and (1) that SDM stands in stark contrast to single-stage DM, wherein downstream consequences are not necessarily considered. After all, a core challenge of SDM is to optimize the cumulative outcome over the various multi-stage DM points. Practically speaking, Real-World Scenarios (RWS) typically involve SDM amidst ever-evolving environs; along this vein, RWS SDM often occurs amidst varying degrees of A/U, whose proposed handling, among other approaches, is discussed in the following Section IB.

B. Tackling SDM Amidst A/U

At the core, RWS AIS SDM A/U (AIS-SDM-A/U) tends to involve several steps/DM points/DMIPs amidst a complex and dynamic ecosystem. Apart from the transportation-related venues previously described in Section IA, AIS-SDM-A/U has also become more prevalent in the energy sector (e.g., for the positing of energy demand over a certain period of time and optimizing the involved power grid to handle the associated consumption), the financial sector (e.g., for the determination of the allocation and re-allocation of financial assets to optimize Return on Investment or ROI), the healthcare sector with regards to patient monitoring, diagnosis, and treatment planning (e.g., for the selection of a series of treatments to attain, ultimately, a more optimal outcome), etc. Apart from these referenced sectors, AIS-SDM-A/U is becoming prevalent in other sectors as well. Yet, at its core, SDM is a formidable area to tackle with the well-known Bellman notion of the "curse of dimensionality" (wherein contending with an increasing number of features can segue to an exponential increase in computational cost and/or a prospective marked decline in performance), and *the question of how to handle A/U remains an ongoing challenge* [4]; along this vein, features, dimensions, and criteria should be differentiated. Features typically equate to what the involved entity "is" or "does" (i.e., property/characteristic, functionality), dimensions relate to how the entity can be delineated or classified/categorized, and criteria include the standards and/or benchmarks utilized to assess the entity. Given these distinctions and the associated backdrop, certain *preset* approaches, such as that of a Lower Ambiguity, Higher Uncertainty (LAHU) and Higher Ambiguity, Lower Uncertainty (HALU) Module (LHM), have been advanced in an effort to tackle the AIS-SDM-A/U challenge. The functions of the LAHU and HALU LHM components are described in Table II.

TABLE II. LAHU/HALU MODULE (LHM)

LHM Components	***Definitions***
Lower Ambiguity, Higher Uncertainty (LAHU)	"Under a Compressed Decision Cycles (CDC) 'paradigm or tight time constraints,' the LHM 'accepts higher uncertainty (i.e., sparse data) given the condition of lower ambiguity' (i.e., Lower Ambiguity, Higher Uncertainty or LAHU)," "and this roughly translates to the consideration that an isomorphic scenario has manifested previously within the available historical data" [5], [6], and those Digital Object Identifiers (DOIs) of Table IV.
Higher Ambiguity, Lower Uncertainty (HALU)	Under a paradigm of Uncompressed Decision Cycles (UDC), meaning, "if there exists a condition of Higher Ambiguity" conjoined with Lower Uncertainty (i.e., HALU), "wherein the isomorphic scenario is nonexistent within the historical data, there will be a proactive seeking of more data 'to lower uncertainty' so as to move towards a more acceptable state" [5], [6], and those DOIs of Table IV.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

With regards to the LHM's repertoire of experience, this can include the various data categories of Table III: *Non-Operational Data*, *Situational Awareness Data*, and *Operational Data,* among others. These comprise, among other facets, the "repertoire of experience" (e.g., the encountering of similar scenarios) of the LHM.

TABLE III. LHM DATA CATEGORIES

Data Categories	***Definitions***
Non-Operational Data (NOD)	"historical and forensic data that has been ingested, apriori" or near-contemporaneously, "to serve as a baseline" for contextualizing the *Operational Data*" [first DOI of Table IV].
Situational Awareness Data (SAD)	environs data that is "contextualized and integrated with *Operational Data*" for the purpose of appraising *Operational Data* "prior to the exigency circumstance;" "in this way…lessons can be learned and leveraged without necessarily needing the immediate performance required of *Operational Data*" [first DOI of Table IV].
Operational Data (OD)	"this data is, when contextualized by *Situational Awareness Data* and *Non-Operational Data*, can be quite indicative and better lend to the immediate performance expected of quasi-real-time data for decision-making" [first DOI of Table IV].

To recap Tables II and III, and as discussed in the first DOI of Table IV, under a LAHU paradigm, given a sufficient repertoire of experience, the tolerance for uncertainty is raised, thereby lowering "the need to turn *Non-Operational Data* and *Situational Awareness Data* into immediate performance data" [first DOI of Table IV]. Suffice it to say, for future circumstances, wherein immediate DM is necessary, the now formulated set of *Deep Belief Heuristics* may be utilized. The practical utility of the LHM (i.e., its enablement) is delineated in [5], [6], and those DOIs of Table IV below.

TABLE IV. EXEMPLAR DIGITAL OBJECT IDENTIFIERS (DOIS) PERTAINING TO THE NOTIONS OF LAHU/HALU

Facet	***DOI***
LAHU; HALU	• 10.1109/IEMCON.2019.8936241 • 10.1109/IAICT62357.2024.10617473 • 10.1109/GEM61861.2024.10585580 • http://dx.doi.org/10.2139/ssrn.5183492 • http://dx.doi.org/10.2139/ssrn.4984663 (DOIs are generally not assigned to patents, such as [5] and [6]).

As described by the last DOI of Table IV, central to the LHM are various supporting modules, such as that of an Isomorphic Paradigm (IsoP) Comparator Similarity Measure (ICSM) Module (ICSM2) as well as a Metaheuristic Algorithm Module (MAM), among other modules. These are briefly described as follows, and further clarification will be provided in Section II.

1) LHM's ICSM2

The ICSM's decided actions may be predicated upon whether the compared paradigm actually needs to undergo a computationally more extensive IsoP examination. For example, in some cases, such as for unordered sets, the ordering of the edges (and their weights) may not necessarily be relevant, as only the nodes (and their values) (hereinafter, "N+V") need to be compared; in other cases, such as for ordered sets, the edges and the sequencing of the nodes is of significance. Exemplar ICSM2 considerations are presented, in somewhat logical order, within Table V below.

TABLE V. EXEMPLAR CONSIDERATIONS PRIOR TO AND INCLUDING THAT OF ISOMORPHIC COMPARISON

ICSM2 Considerations	***Definition***
Unordered Set (UnS)	Set of disparate constituents, wherein the order of the constituents is not relevant. By way of example, {1, 2, 3, 4, 5} equates to {5, 3, 1, 4, 2}.
Equal Sets (EqS)	A pair of sets S and S' is equal if and only if (iff) each constituent of S is also a constituent of S'; moreover, the order of the constituents is not relevant. By way of example, if S = [1, 2, 3, 8, 9, 10] and S' = {9, 3, 1, 2, 10, 8}, then S=S'.
Equivalent Sets (EquivS)	A pair of sets S and S' is considered equivalent if the number of constituents in S and S' is the same (i.e., same cardinality). By way of example, if S = {1, 3, 5, 7, 9} and S' = {2, 4, 6, 8, 10}, then S and S' are considered to be equivalent.
Ordered Set (OrS)	Set of disparate constituents, wherein the order of the constituents is relevant, and the constituents can be ordered and compared via operators, such as <. By way of example, an ordered set might be {1, 2, 3, 5, 6, 8, 9, 10}, whereas an unordered set might be {6, 5, 1, 2, 3, 10, 9, 8}.
Partially Ordered Set (POSET)	Set of disparate constituents, wherein the constituents might or might not be able to be ordered and compared, since operators such as <= can yield different variations. By way of example, Calcworkshop (https://calcworkshop.com/relations/partial-order/) provides some examples, which we extrapolate upon in the way of {a < b < c < d <= e <= f }, {a <= b <= c < d < e <= f < g}, and {a < b < c <= d <= e < f}, which are shown diagrammatically in Table VI.
Unordered Sets with Isomorphism (UnS-Iso)	For a set of disparate constituents, wherein the constituents are unordered, if there is a one-to-one relationship (i.e., bijection), then the unordered sets are likely isomorphic. By way of example, if S={1, 2, 3, 4, 5}, S'={a, b, c, d, e}, and 1<->a, 2<->b, 3<->c, 4<->d, and 5<->e (wherein each constituent in S relates to a unique constituent in S'), then S and S' are considered to be isomorphic.
POSETs with Isomorphism (POSET-Iso)	For a set of disparate constituents, wherein the constituents are considered to be within a POSET, if there is a bijection, then the POSETs are likely isomorphic. By way of example, if S={S_1, S_2, S_3, S_4, S_5}, S'={S'_2, S'_4, S'_5, S'_3, S'_1}, and S_1<->S'_2, S_2<->S'_5, S_3<->S'_1, S_4<->S'_4, and S_5<->S'_3, (wherein each constituent in S relates to a unique constituent in S'), then S and S' are considered to be isomorphic. To demonstrate this, online tools are available, such as https://graphonline.top/en/?graph=xPLjwOkrglDRgYeS,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

	among others.
Isomorphism Variants (IV)	Extrapolating upon the POSETs with isomorphism, there are also permutations that are actually isomorphism variants (e.g., automorphism, which is a particular type of isomorphism that has a symmetrical structure), which Lemons nicely casts and for which examples are shown in Table VII as IV#1 through #3 [7].

TABLE VI. EXEMPLARS OF ISOMORPHIC PARTIALLY-ORDERED SETS (POSETS)

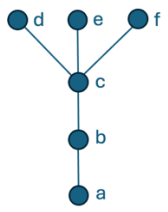
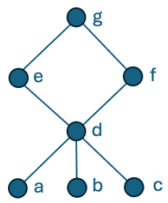
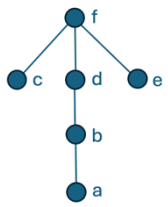
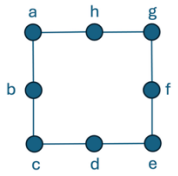
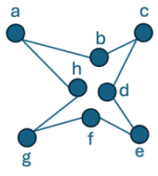
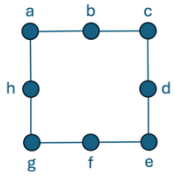
POSET #1	***POSET #2***	***POSET #3***
		
Maximal: d, e, f	Maximal: g	Maximal: f
Greatest: none	Greatest: g	Greatest: f
Minimal: a	Minimal: a, b, c	Minimal: a
Least: a	Least: none	Least: a

TABLE VII. EXEMPLARS OF ISOMORPHIC VARIANTS (IVS)

IV #1	***IV #2***	***IV #3 (automorphism)***
		

The LHM ICSM2's considered pathways for undertaking the requisite pre- and isomorphic/non-isomorphic comparisons, such as alluded to by Table V, are underpinned by the LHM's MAM. After all, depending upon the time available (as was depicted in Figure 1), certain pathways may likely be more computationally tractable. The notion of Criteria Weights (CW), such as that of an Adaptive Criteria Weighting System (ACWS), is crucial in this regard, and exemplar prior experimentation is listed in Table VIII; it should be remembered that CW/ACWS is pivotal for Multi-Criteria Decision-Making (MCDM) and that the notion of the CW/ACWS dovetails with the Interstitial bespoke (b) Shapley Additive Explanations (SHAP)-Owen Amalgam or Interstitial b-SHAP-Owen Amalgam (IbSOA), which designates an importance value to each feature; after all, CW/ACWS systematically assesses and ranks alternatives or features based upon their relative importance (as determined by the CW). IbSOA will be further addressed in Section IIID.

TABLE VIII. EXEMPLAR DOIS PERTAINING TO CRITERIA WEIGHTING EXPERIMENTATION

Facet	***DOI***
CW, ACWS	• 10.1109/ICSGTEIS60500.2023.10424230 • 10.1109/ICPEA56918.2023.10093212 • 10.1109/AIIoT61789.2024.10579033 • http://dx.doi.org/10.2139/ssrn.4984663 • http://dx.doi.org/10.2139/ssrn.5183492 • 10.1109/AIIoT65859.2025.11105315

2) *LHM's MAM*

Proceeding from the ICSM2 discussion to that of the MAM, the MAM's decided actions, particularly amidst the temporal limitations of CDC, may also be predicated upon what type of Decision-Making Problem (DMP) is encountered. In the case of SDM, each DM point may be handled by whether the DMP is any of, among others, the following: (1) Programmed DMP, (2) Analytical DMP, and (3) Non-Programmed DMP. The definitions for these are shown in Table IX below.

TABLE IX. TYPES OF DECISION-MAKING PROBLEMS (DMPS)

Type of DMP	***Definition***
Programmed DMP (PDMP)	Prototypical Structured Problems (PSP), as pertains to DMP, with well-established Standard Operating Procedures (SOPs) and approach vectors.
Analytical DMP (ADMP)	Mid-range Semi-Structured Problems (MRSSP), as pertains to DMP, that require more in-depth analysis and a more comprehensive approach; it may also require decomposing the DMP into smaller, more manageable (i.e., computationally tractable) DMPs.
Non-Programmed DMP (NPDMP)	More Complex Unstructured Problems (CUP), as pertains to DMP, that may require a bespoke approach.

As the MAM, by way of example, considers each DMP, the insights provided by the IbSOA will be invaluable, as the likely behavior within the MNTZ may be pivotal in ensuring that a prudent pathway, within the given time constraints, is chosen. This will be further expounded upon in Sections IIC and IIID. In the context of the MAM, as noted in the 6th DOI of Table VIII, according to Hjeij, in the journal paper, "A brief history of heuristics: how did research on heuristics evolve," "the use of a heuristic is 'inevitable where no method to find an optimal solution exists or is known,' particularly when 'the problem and/or the optimality criterion is ill-defined'." To add to this, Bobadilla-Suarez, in the journal paper, "Fast or Frugal, but Not Both: Decision Heuristics Under Time Pressure," notes "that heuristics do not form a uniform class and that "more frugal heuristics will not necessarily be faster to implement than less frugal ones. Similarly, less frugal strategies can be fast given the right stimuli [i.e., accelerants]" [6th DOI of Table VIII]. Drake as well as Fisher & Thompson add to what Bobadilla-Suarez asserted; as delineated in the journal paper, "A Novel Cooperative Multi-State Hyper-Heuristic

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

for Combination Optimization Problems," "'mixing and combining different Low-Level Heuristics [LLH] produced better quality solutions than if they were applied separately' and 'showed that individual heuristics may be particularly effective at certain stages…but perform poorly at others,' as each individual heuristic may involve particular methods (as Zayat and Watrobski well noted) [6th DOI of Table VIII]. In essence, 6th DOI of Table VIII features the case study, wherein it "might be prudent to avoid utilizing AHP, SAW, and [Preference Ranking Organization Method for Enrichment Evaluation] PROMETHEE when handling negative values" [6th DOI of Table VIII].

The presets of ICSM2 and MAM (two of the key components of LHM) have now been touched upon. Prior to the unpacking of the proposed third component of LHM, CAIR-A, some background information will be provided. For the reader's convenience, a listing of the acronyms utilized thus far as well as for the sections that follow is being provided in Table X below.

TABLE X. LISTING OF ACRONYMS UTILIZED

Acronym	***Full Form***
A	Ambiguity
ACWS	Adaptive Criteria Weighting System
ADMP	Analytical DMP
AI	Artificial Intelligence
AICDS	Artificial Intelligence Control and Decision System
AIS	Artificial Intelligence System
AS	Autonomous System
ADMB	Automatic Differentiation Model Builder
ASR	Automatic Speech Recognition
A/U	Ambiguity/Uncertainty
AUV	Autonomous Underwater Vehicle
b-SHAP	Bespoke Shapley Additive Explanation
b-SHAP-Owen	Bespoke Shapley Additive Explanation Owen
c-SHAP	classical Shapley Additive Explanation
CAA	Conversational Artificial Intelligence Agent
CAIR	Conversational Artificial Intelligence Robustness
CAIR-A	Conversational Artificial Intelligence Robustness Accelerant
CAP	Component Assignment Problem
CBR	Case-Based Reasoning
CDC	Compressed Decision Cycle
CDS	Control and Decision System
CPRLD	**C**onstriction Factor-**P**article Swarm Optimization-**R**obust Convex Relaxation-**L**ong Short-Term Memory-**D**eep Convolutional Neural Network
CF	Constriction Factor
CRITIC	CRiteria Importance through Intercriteria Correlation
CWS	Criteria Weighting System
CWT	Continuous Wavelet Transforms
CVA	Conversational Virtual Agent
CW	Criteria Weight
DCGAN	Deep Learning Generative Adversarial Network
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
DM	Decision-Making; Decision Maker
DMP	Decision-Making Problem
DMIP	Decision-Making Inflection Point
DOI	Digital Object Identifier
E	Execution
ELECTRE	ÉLimination Et Choix Traduisant la REalité
EM	Estimation-based Method
EUT	Expected Utility Theory
FCSO	Finite-Change Shapley-Owen
FN	Fuzzy Number
FS	Fuzzy Set
GAN	Generative Adversarial Network
GBR	Graph-Based Reasoning
G*f*FN	Generalized *f-sided* Fuzzy Number
GFN	Generalized Fuzzy Number
GFS	Generalized Fuzzy Set
G*Hp*FN	Generalized *Heptagonal* Fuzzy Number
G*Hx*FN	Generalized *Hexagonal* Fuzzy Number
GI	Graph Isomorphism
GL*f*FN	Generalized Linear *f-sided* Fuzzy Number
GLIV*Pe*FN)	Generalized Linear Interval-Valued *Pentagonal* Fuzzy Number
Glocal	a portmanteau "global" and "local"
GLIV*Pe*FN	Generalized Linear Interval-Valued *Pentagonal* Fuzzy Number
GL*Pe*FN	Generalized Linear *Pentagonal* Fuzzy Number
GN	Grey Number
G*N*FN	Generalized *N-sided* Fuzzy Number
GNL*f*FN	Generalized Non-Linear *f-sided* Fuzzy Number
GNLIV*Pe*FN	Generalized Non-Linear Interval-Valued *Pentagonal* Fuzzy Number
GNL*Pe*FN	Generalized Non-Linear *Pentagonal* Fuzzy Number
G*Oc*FN	Generalized *Octagonal* Fuzzy Number
G*Pe*FN	Generalized *Pentagonal* Fuzzy Number
GPL	General Public License
GPU	Graphics Processing Unit
GS	Grey Set
GSO	Generic Shapley-Owen
G*Tp*FN	Generalized *Trapezoidal* Fuzzy Number
G*Tr*FN	Generalized *Triangular* Fuzzy Number
HALU	Higher Ambiguity, Lower Uncertainty
HH	Hyper-Heuristic
I	Interpretability
IA	Independence Axiom
IBE	Inference to the Best Explanation
ICSM	Isomorphic Paradigm Comparator Similarity Measure
ICSM2	Isomorphic Paradigm Comparator Similarity Measure Module
IFF	If and Only If
IFN	Intuitionistic Fuzzy Number
IFS	Intuitionistic Fuzzy Set
IM	Intelligent-based Method
IPOPT	Interior Point OPTimizer
IsoP	Isomorphic Paradigm
IUC	Inherent Uncertainty Construct
IVFN	Interval-Valued Fuzzy Number
IVFS	Interval-Valued Fuzzy Set
IVIFS	Interval-Valued Intuitionistic Fuzzy Set
LAHU	Lower Ambiguity, Higher Uncertainty
LLH	Lower-Level Heuristic
LHM	**L**ower Ambiguity, Higher Uncertainty/**H**igher Ambiguity, Lower Uncertainty **M**odule
LLM	Large Language Model
LOE	Line of Effort
LSTM	Long Short-Term Memory
MADM	Multi-Attribute Decision-Making
MADM/MODM	Multi-Attribute Decision-Making/Multi-

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

SM/OM	Objective Decision-Making Subjective Method/ Objective Method
MAM	Metaheuristic Algorithm Module
MCDM	Multi-Criteria Decision-Making
M_i	Imprecision Membership
MILP	Mixed Integer Linear Programming
MINLP	Mixed Integer Non-Linear Programming
MIP	Mixed Integer Programming
ML	Machine Learning
ML2	Machine Learning on Machine Learning
M_m	Median Membership
MMSO	**M**ulti-Attribute Decision-Making/**M**ulti-Objective Decision-Making **S**ubjective Method/ **O**bjective Method
MNTZ	Monotonic/Non-Monotonic Transition Zone
MODM	Multi-Objective Decision-Making
MR	Monotonic Reasoning
NDZ	Non-Detection Zone
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NMR	Non-Monotonic Reasoning
NOD	Non-Operational Data
NPDMP	Non-Programmed DMP
OD	Operational Data
OM	Objective Method
OSNS	Optimal Shapley-Nondominated Solution
OSONS	Optimal Shapley-Owen-Nondominated Solution
OUF	Over/Under Frequency
OUV	Over/Under Voltage
P-RM	Primary Reasoning Mechanism
PDMP	Programmed DMP
PJD	Phase Jump Detection
POSET	Partially Ordered Set
PROMETHEE	Preference Ranking Organization Method for Enrichment Evaluation
PSO	Particle Swarm Optimization
r-IsoP	Relaxed Isomorphic Paradigm
RMS	Relationship/Membership Stream
RR	Rank Reversal
RCAP or dP/dt	Rate of Change of Active Power
RCR	Robust Convex Relaxation
RDM	Robust Dialogue Management
RM	Reasoning Mechanism
ROCOF of df/dt	Rate of Change of Frequency
ROCPAD	Rate of Change of Phase Angle Difference
ROYG	Red-Orange-Yellow-Green
ROI	Return on Investment
RTIA or R-III	Reverse Triple I Algorithm
RWS	Real-World Scenario
S	Sensitivity
S-RM	Secondary Reasoning Mechanism
SAD	Situational Awareness Data
SAE	Society of Automotive Engineers
SCSO	"Squared Cohorts" Shapley-Owen
SDM	Sequential Decision-Making
SDP	Semi-Definite Programming
SQP	Sequential Quadratic Programming
SHAP	Shapley Additive Explanation
SNOPT	Sparse Nonlinear OPTimizer
SIS	Subsethood Inference Subsethood
SM	Subjective Method
SOP	Standard Operating Procedure (SOP)
STEA	System Transparency Explainability & Accountability
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution

TPU	Tensor Processing Unit
TIA	Triple I Algorithm
T1FS	Type 1 Fuzzy Set
T2FS	Type 2 Fuzzy Set
U	Uncertainty; [performance under] Uncertainty
UAV	Unmanned Aerial Vehicle
UDC	Uncompressed Decision Cycle
USV	Unmanned Surface Vehicle
V	Validity
VAP	Value-added Proposition
VBSO	Variance-Based Shapley-Owen
VIKOR	VlseKriterijumska Optimizacija I Kompromisno Resenje
vTHD	Voltage Imbalance and Total Harmonic Detection
VVS	Voltage Vector Shift
WM	Wavelet Transform-based Method

With these acronyms and their full forms in hand, it seems prudent to provide a quasi-dependencies equation so as to facilitate the parsing of this paper. The first equation will take the form presented in (1a), and the standard convention of a dashed arrow (wherein the dependent element is to the left of the tail, and the element that it is dependent upon is to the right of the arrowhead) is utilized.

AIS⇢**S**DM⇢**L**HM⇢ICSM2+MAM (1a)

The extended version for constituents supporting the module for LHM⇢ICSM+MAMF, for the purposes of this paper, becomes (1b).

LHM⇢**I**CSM2+**M**AM+**C**AIR-A (1b)

As it is understood that the LHM instantiation now includes CAIR-A as well, such as is reflected in (1b), (1) will simply be referred to as ASL-IMC. Proceeding on, the second equation will take the form, as presented in (2), using the same convention.

RDM ⇢**M**R/NMR⇢**M**NTZ⇢**I**bSOA (2)

For convenience, (2) will be referred to as RMMI. The follow-on third equation will then take the form, as is presented in (3), using the same convention.

ASL-IMC⇢RMMI (3)

For this paper, (3) will be the overarching operative quasi-dependencies equation, and (3) will be unpacked within this paper in the following fashion. Section I had provided an overview of the SDM challenge and provided some exemplar *presets* (e.g., LHM, ICSM2, MAM) as well as introduced a new module (e.g. CAIR-A) to be utilized for the experimentation herein; Section I also addressed (1a). The remainder of this paper is organized as follows. Section II presents the backdrop/background for an AIS approach

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

towards DMP, such as SDM, which is predicated upon a robust harmonizing/counterpoising of MR and NMR; in turn, a successful MR/NMR counterpoising is dependent upon robust insights into the MNTZs. The MNTZs, in turn, are informed by IbSOA. Then, the previously referenced bespoke mechanism in the form of the referenced CAIR-A Module, which can assist with the ongoing CAA challenge of Robust Dialogue Management (RDM), is touched upon. Section III lays out some theoretical foundations and picks back up on the discussion regarding IbSOA and its derived insights, which can be instrumental at the MNTZs by gleaning the interplays among local, glocal (a portmanteau "global" and "local"), and global, particularly when conjoined with a robust MADM/MODM SM/OM (MMSO) mechanism; the section then proceeds with presenting some precursor experimentation as well as an interim discussion on IbSOA. Section IV provides some further thoughts as well as concluding remarks, and proposed future work closes the paper.

II. BACKGROUND

The AIS approach towards DMP, such as SDM, may involve a variety of Reasoning Mechanisms (RMs). In the context of the referenced LHM, Analogical Reasoning, Abductive Reasoning, Inductive Reasoning, Probabilistic Reasoning, and Temporal Reasoning, among others, may be utilized. These primary RMs are described in Table XI.

TABLE XI. TYPES OF PRIMARY RMS (P-RMS)

Type of Primary RM	*Definition*
Analogical Reasoning (AnaR)	Firt put it quite nicely and simply by saying that analogical reasoning is a "type of thinking that relies upon an analogy" [8]. Gentner affirms and deems it to be "the ability to perceive and use relational similarity between two situations or events" [9]. In a similar fashion, Thibaut refers to it as a "process in which a base domain and a target domain are compared in order to find relational correspondences" [10].
Probabilistic Reasoning (ProbR)	Nandi opines that probabilistic reasoning "is a framework used to make inferences and decisions under uncertainty" [11]. Nafar concurs, cites Oaksford and Chater, and notes that "probabilistic reasoning often aligns with Bayesian Rationalism," and within an AI context, Nafar refers to it as a "mapping [of] uncertainty…to a Bayesian probabilistic framework" [12].
Temporal Reasoning (TempR)	Leeuwenberg posits that temporal reasoning is "the process of combining different temporal cues into a coherent temporal view" [13]. Xiong unpacks this further and asserts that some of the types of temporal reasoning include "sequencing, duration, frequency, simultaneity, temporal relation, comparative analysis and facts extraction," which can be construed roughly as determining chronological order, how long an event lasted, how much time elapsed between the start of one event and the start of another, when a particular event occurred, whether events happened at the same time, and contrasting multiple events, respectively [14].
Deductive Reasoning (DedR)	Grote-Garcia put it quite well by noting that "deductive reasoning is the process of using general premises to draw specific conclusions" [15]. Along this vein, Taylor & Francis recaps this by stating that deductive reasoning "involves moving from the general to the specific, and it is used to draw conclusions based on known facts or assumptions" [16].
Inductive Reasoning (IndR)	The University of Illinois Springfield puts it succinctly: "inductive reasoning is the ability to combine pieces of information that may seem unrelated to form general rules or relationships" [17]. In essence, inductive reasoning is a technique of deriving conclusions by progressing from the specific to the general.
Abductive Reasoning (AbdR)	Sandoval-Hernandez asserts that "abductive reasoning involves starting with surprising or puzzling observations and inferring the most likely explanations" [18]. Thagard views it in a slightly more subdued fashion, wherein "explanatory hypotheses are formed and evaluated" [19]. Belzen follows Thagard and asserts that "abductive reasoning describes the theory-based attempt of explaining a phenomenon by a cause" [20].

In turn, both Analogical Reasoning and Abductive Reasoning may leverage Case-Based Reasoning (CBR), while the former may also utilize Graph-Based Reasoning (GBR). These are described in Table XII below.

TABLE XII. TYPES OF SECONDARY RMS (S-RMS)

Type of RM	*Definition*
Case-Based Reasoning (CBR)	Yan describes CBR as being "based on the cognitive assumption that similar problems have similar solutions" [21]. Taylor & Francis captures this as CBR being "a problem-solving approach that involves using past successful solutions to similar problems to solve new problems" [22].
Graph-Based Reasoning (GBR)	GBR approaches have been utilized for the purposes of, among others, that described in [23]. In essence, it is a technique that buttresses reasoning capabilities by characterizing problems as graphs. Cao adds an interesting addendum to this notion by putting forth the "Reasoning Graph Verifier" (a.k.a., "GraphReason") to "analyze and verify the solutions generated by LLMs," which Cao asserts "enhances the reasoning abilities of LLMs" and "also outperforms existing verifier methods in terms of improving these models' reasoning performance" [24];

Interestingly, GBR is often applied by AIS for Fault Detection (discerning anomalies and positing causes/root causes), Diagnosis, and NLP/NLU (for positing the context and intended meaning), such as described by the papers associated with the DOIs of Table XIII.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE XIII. EXEMPLAR DOIS PERTAINING TO THE NOTIONS OF DETECTION, DIAGNOSIS, NLP/NLU

Facet	*DOI*
Detection	• 10.1109/CMD48350.2020.9287173 • 10.1109/CMD48350.2020.9287173 • 10.1109/CMD48350.2020.9287281 • 10.1109/CMD48350.2020.9287262 • 10.1109/CMD48350.2020.9287299
Diagnosis	• 10.1109/ICOIACT46704.2019.8938444
NLP/NLU	• https://ssrn.com/abstract=3789767

A. AIS for Various Sectors and a Proxy Application

AIS approaches, particularly for AI Control and Decision Systems (CDS) (AICDS), as pertains to DMP/SDM, remain an active research area not only for the sector touched upon in Section IA, but also for the sectors of, among others, healthcare, finance, and energy. These sectors tend to involve elongated temporal spans for evaluation metrics, and definitive results may be challenging to come by. While the referenced AUVs, USVs, UAVs, and self-driving cars can indeed be somewhat comprehensively evaluated, the benchmarking thereof can be cost prohibitive and may require elongated temporal spans (e.g., testing them under various weather conditions, such as detection capability against the white background of snow). Accordingly, the analysis of Abductive Reasoning and benchmarking of NLU as well as NLG can, potentially, be more cost effective and economize on time when using a proxy application, such as Conversational AI (wherein AI is used to mimic human-like conversational dialogue). For a CAA (a.k.a., Conversational Virtual Agent or CVA), beyond the essential first step of ASR, NLU and NLG are formidable downstream steps. It is imperative that the AIS be able to discern the conversational context as well as the overarching intent of the user (i.e., NLU) and respond accordingly (i.e., NLG). Presuming that NLU and meaningful NLG can be satisfactorily achieved, RDM becomes central. In essence, RDM involves the CAA being able to carry on a coherent conversation, wherein validity is, ideally, sustained both within the single interaction as well as over the course of several interactions (a.k.a., multi-turn conversations). Even within a single interaction, the conversation may be quite nuanced with a variety of inflection points. These inflections may also occur over the course of the multi-turn conversation. Maintaining consistency, coherency, and validity throughout these multi-turn conversations is central for meaningful/productive dialogues (i.e., RDM). Significantly, consistent validity inspires trust in the CAA. As the time-dependent priorities of the user and information being consumed by the CAA may dynamically change, the varying context could contradict prior information. Therefore, it is critical that the CAA has the capability for the graceful handling of contradictions with due care so as to help ensure conversational fluidity and maintain ongoing confidence and trust in the CAA. This capability is rooted in the successful counterpoising of the CAA's Monotonic Reasoning (MR) and Non-Monotonic Reasoning (NMR) and the handling of MR/NMR and the MNTZs (a.k.a., CAIR-A). MR/NMR (and the MNTZs) will be further unpacked in the following Sections IIB through IIC.

B. MR and NMR Counterpoising

Against the backdrop of AI and RMs, MR and NMR are two of the major pillars for AI-centric logic/RMs. These are described in Table XIV below.

TABLE XIV. TYPES OF AI-CENTRIC LOGIC/RMS

Type of AI-centric logic/RMs	**Definition**
Monotonic Reasoning (MR)	MR-centric responses will remain consistent throughout time despite whatever new information might arrive.
Non-Monotonic Reasoning (NMR)	NMR-centric responses allow for modification and/or retraction of prior assertions.

As noted in Section IA, "as the priorities of the user and arriving information may alter the context and/or contradict prior information," the harmonizing and counterpoising of the MR and NMR becomes critical for maintaining logical consistency and interconnectedness among the multi-turn conversations (i.e., coherence); if the constituent elements of the multi-turn conversation are indeed logically related, then the overall dialogue should be relatively free of contradictions. This constitutes "conversational coherence." There are a variety of evaluation tools in this regard. For example, the "ConversationCoherence Evaluator" is purported to be "a tool designed to check the coherence of conversations by an AI," as "it evaluates whether each response in a conversation logically follows from the previous messages, ensuring that the AI maintains context and relevance throughout the interaction" [25]. However, "conversational coherence" is, actually, quite difficult to maintain in RWS, where information is often sparse/incomplete and/or ambiguous/uncertain. Depending upon "what" and "when" the information is made available (the issues of provenance/pedigree regarding the "who" and the "where" will not be discussed here; rather, they will be discussed in future work), a specific RM may be used, such as shown in Table XV.

TABLE XV. TYPES OF PRIMARY REASONING MECHANISMS (P-RMS)

Information Available: "What"/"When"		***Type of RM Utilized***	***Conclusion***
• Facts • Accepted Truths • Rules • Scientific Laws • Mathematical Theorems • Established Principles • Specific	Used in an iterative fashion	Deductive Reasoning	Guaranteed to be true, if the premises and argument are valid.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Information • Logical Connections			
Starts with the same set as Deductive Reasoning (if available), but also involves Probabilistic approaches	Typically front-loaded, but can also "unfold in a bottom-up" fashion	Inductive Reasoning	Likely to be true, but it could be false despite the observations being accurate.
Starts with the same set as Deductive Reasoning (if available), but also involves hypotheses, assessments, and best-fit approximations	Typically front-loaded, but as it is trend-sensitive as well as aberration-sensitive, it can also "unfold in a bottom-up" fashion	Abductive Reasoning	Plausible best guess approximation or a posit as to the optimal explanation; along this particular vein, Harman is well-known for his research involving "Inference to the Best Explanation" (IBE) [26].

With regards to the various RMs presented in Table XV, first, deductive reasoning will be addressed. Ferguson proclaims that "classic deductive logic entails that once a conclusion is sustained by a valid argument, the argument can never be invalidated, no matter how many new premises are added. This derived property of deductive reasoning is known as *monotonicity*" [27]. Bundy and Wallen restates this as "the *monotonicity* of deductive logic," wherein "the addition of new axioms to a set of axioms can never decrease the set of theorems or facts" [28]. Fuhrmann summarizes the aforementioned with the notion that "deductive inference, at least according to the canons of classical logic, is *monotonic*; if a conclusion is reached on the basis of a certain set of premises, then that conclusion still holds if more premises are added" [29].

Next, turning to abductive reasoning, Hentenryck asserts that "as a form of reasoning appropriate for handling incomplete information, abduction is also closely related to non-monotonic reasoning" [30]. Paul affirms this by clearly stating: "abduction is a form of non-monotonic reasoning" [31]. Lagerkvist follows on by summarizing as follows: "one of the best-known examples of non-monotonic reasoning is *abductive* reasoning" [32].

Proceeding along to inductive reasoning, Leidinger cites Hans with regards to testing "nonmonotonic reasoning among other inductive reasoning tasks" [33]. What is particularly interesting is that Hans notes that in his testing, Large Multimodal Models (LMMs)/LLMs (e.g., GPT-3.5, GPT-4) were not able to capture "human behavior on the non-monotonicity phenomena," and actually took "the opposite" position [34]. Along this vein, although GPT-4 was much more successful in capturing human behavior than GPT-3.5, Hans highlighted the "notable exception" in "its failure to capture the phenomenon of premise non-monotonicity" with regards to inductive reasoning (a.k.a., property induction) [34]. *Hence, the handling of non-monotonic reasoning remains an ongoing challenge that may be impacting the performance of the newer LMMs/LLMs being released*. Generally speaking, Kazemi notes that LLMs have difficulty contending with contradictory information (thereby segueing into the challenges of contending with the non-monotonic realm). [35].

In brief, deductive reasoning is often construed as a form of MR (wherein the addition of relevant information buttresses conclusions reached based upon recitals of fact and evidentiary material) while abductive reasoning is often considered to be a form of NMR (wherein new information can potentially reverse [i.e., cause to retract] prior inferences reached through evidentiary material and reasoning). Hence, inductive reasoning is typically taken to be non-monotonic, but the literature has also noted cases where it is weakly monotonic; for example, Janke delineates how NMR "is inherently required in several approaches to inductive inference" [36]. Overall, insights into MR/NMR behavior (within the MNTZ) and an apropos harmonizing/counterpoising of MR/NMR can provide the requisite flexibility for RDM and the desired ensuing CAA conversational coherence.

C. *Insights into the MNTZ*

As noted at the end of Section IIB, the successful counterpoising of MR/NMR necessitates meaningful insights into the MNTZ. Yet, as discussed in [2], the discerning/comprehending of the behavior at the MNTZ can be quite challenging. In addition, [2] noted that "maintaining coherence and monotonicity is non-trivial, as the involved AIS might discern connections (particularly those that are non-monotonic) within the evolving dataset. In the context of CAIR at-large (CAIR-A, specifically), non-monotonic aspects can arise as incoming information can re-contextualize and/or contradict matters. Yet, enforcing a strict monotonic paradigm can segue to an unnatural rigidity and/or incorrect/irrelevant responses by the" CAA. Hence, the counterpoising of MR/NMR is non-trivial.

As further noted in [2], "the shift of the involved variables from a monotonic to a non-monotonic paradigm can be quite unexpected and occur more frequently than anticipated/desired." For AIS-SDM-A/U, the spawning of "non-monotonic, non-polynomial, and even non-continuous functions" does indeed occur more frequently than expected or desired [37]. Restated, as pertains to the MNTZ, there exists a proclivity for "spawning to the NP-Hard side," and this is similar to the situation, "wherein the transformation of non-convex Mixed Integer Non-Linear Programming (MINLP) to convex problems often spawn further non-convex MINLP problems" [2]. With regards to

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Monotonic/Non-Monotonic and Linear/Non-linear (wherein Non-Monotonic can be Continuous or Discontinuous and Non-Linear can be Polynomial or Non-Polynomial), Figure 2 conveys some of the prospective pathways to convex form (in green) as well as the pathways that remain nonconvex (in red) [2]; MIP equates to Mixed Integer Programming, and MILP equates to Mixed Integer Linear Programming.

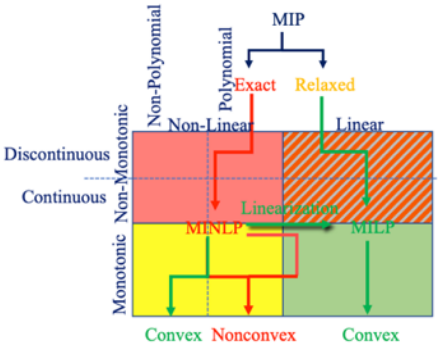


Figure 2. "Non-convex to convex transformation pathways (e.g., non-convex non-monotonic, non-polynomial, non-continuous MINLPs to convex form)" [2].

For the convex form, a myriad of Semi-Definite Programming (SDP) solvers can be leveraged to handle the computations in polynomial time ("presuming further spawning does not occur," since "NP-hard-related spawn can potentially congest matters with an indefinite impasse") [2]. There are a variety of mechanisms that can be useful in this regard, and some of these facets are discussed in Section IID.

D. *Formulating a Bespoke Mechanism for the MNTZ*

As noted in Sections IIA and IIB, the type of reasoning and prospective RM combinatorial (Analogical, Probabilistic, Temporal, Deductive, Inductive, Abductive, etc.) can indeed impact the MR/NMR amalgam and the likely behavior within the MNTZ. For example, as noted in the prologue to Section II, various forms of primary RM (and various secondary RMs, which they may leverage) are utilized in the context of the referenced LHM. Taking the examples of Tables XI and XII, Table XVI is presented, which delineates the referenced RMs and their MR/NMR categorization.

TABLE XVI. EXAMPLE RMS AND THEIR MR AND/OR NMR CATEGORIZATION

RM	***MR/NMR Categorization***
Analogical Reasoning	*Non-Monotonic*; Kerber notes that "two modi of analogical reasoning" "rely on different forms of relevance knowledge that cause non-monotonicity" [38].
Probabilistic Reasoning	*Monotonic* and *Non-Monotonic*; Liu notes that "it has been found that the ability to handle incomplete information or to perform nonmonotonic reasoning does not exist in some probabilistic reasoning mechanisms" (i.e., these mechanisms can only contend with Monotonic), while Grosof defines "an approach to non-monotonic probabilistic reasoning" [39][40].
Temporal Reasoning	*Monotonic* and *Non-Monotonic*; Baral asserts that temporal reasoning is monotonic much of the time, but amidst new incoming information, the understanding of past, present, and future can alter, and this necessitates "that goals be changed non-monotonically" [41].
Deductive Reasoning	*Monotonic*; As noted earlier in Section IIB, Ferguson notes that Deductive Reasoning is monotonic [27]. However, it should be noted that Hunter makes the caveat that deductive argumentation is non-monotonic [42].
Inductive Reasoning	*Weakly Monotonic* and *Non-Monotonic*; as noted earlier in Section IIB, Jantke notes that NMR "is inherently required in several approaches to inductive inference" [36]. Jantke builds upon this by asserting, "monotonic and non-monotonic reasoning is introduced into inductive inference," discusses a "weakly monotonic inductive inference algorithm," and notes that "consistency and monotonicity can hardly be achieved simultaneously" [37]. The general idea is that a monotonic approach can be used to formulate hypotheses in an incremental fashion, but during this formulation, revisions/retractions may be necessary, and this segues into the area of non-monotonicity. Jantke summarizes with the statement "in the area of inductive inference of total recursive functions monotonicity can rarely be guaranteed" [43].
Abductive Reasoning	*Non-Monotonic*; as noted earlier in Section IIB, Hentenryck and Paul have noted that abductive reasoning is in the realm of non-monotonic reasoning [30][31]. Pereira affirms this [44].
Case-Based Reasoning (CBR)	*Cautiously Monotonic* and *Non-Monotonic*; the premise of CBR is that prior cases are examined to determine a best-fix approximation (i.e., ascertaining the most similar cases) to resolve a current situation. However, new cases may reverse the determinations of prior cases, and new information may result in revision/retraction to an earlier conclusion. However, it should be noted that Paulino-Passos makes the caveat by defining a variation, which is cautiously monotonic [45].
Graph-Based Reasoning (GBR)	*Monotonic* and *Non-Monotonic*; when the addition of edges reinforces the conclusion, the paradigm is monotonic, but when the subtraction/negation of edges or addition of edges that obviate prior paths occurs, it is significant to note that Bochman asserts that GBR-related representations and nonmonotonic inheritance/nonmonotonic reasoning are intricately connected [46].

When Table XVI is sorted by the prevalence of monotonicity to non-monotonicity, the results are as depicted in Table XVII with Red-Orange-Yellow-Green (ROYG) color coding, wherein Monotonic is indicated by

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

green and Non-Monotonic is indicated by red; weakly monotonic is indicated by orange, and cautiously monotonic is indicated by yellow.

TABLE XVII. RM-CENTRIC SORTING OF MR AND NMR BY PREVALENCE

RM	***MR/NMR Categorization***	
Deductive Reasoning (DedR)	*Monotonic*	
Probabilistic Reasoning (ProbR)	*Monotonic*	*Non-Monotonic*
Temporal Reasoning (TempR)	*Monotonic*	*Non-Monotonic*
Graph-Based Reasoning (GBR)	*Monotonic*	*Non-Monotonic*
Case-Based Reasoning (CBR)	*Cautiously Monotonic*	*Non-Monotonic*
Inductive Reasoning (IndR)	*Weakly Monotonic*	*Non-Monotonic*
Analogical Reasoning (AnaR)	*Non-Monotonic*	
Abductive Reasoning (AbdR)	*Non-Monotonic*	

Depending upon the amalgam of RMs and ensuing RM pathways initially selected, such as by the LHM-at-large, the related DM points/DMIPs will likely impact the downstream behavior within the MNTZ in terms of a preponderance of monotonic or non-monotonic behavior. Furthermore, as previously noted in the Abstract, "shifts from a monotonic to a non-monotonic paradigm and vice versa in the MNTZs can occur at a higher than anticipated rate." This is also previously in Section IIC noted: "the shift of the involved variables from a monotonic to a non-monotonic paradigm can be quite unexpected and occur more frequently than anticipated/desired" [2]. Accordingly, given the propensity to migrate towards non-monotonic, the ability to well-handle non-monotonic and somewhat mitigate against the shift from a monotonic to a non-monotonic paradigm seems crucial and prudent. Again, the involved counterpoising is key, as the simpler strategy of reliance (readily segueing into overreliance) upon a non-monotonic paradigm can readily lead to complicated models that are prone to overfitting and, quite possible, consume a not insignificant portion of the computational resources available.

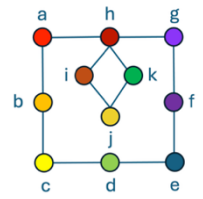
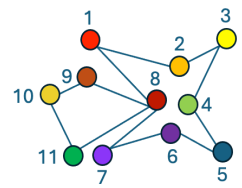
While it seems that focusing on certain RMs (e.g., deductive) might seem computationally more tractable, it would preclude more RWS RM approaches (e.g., ProbR, TempR, and GBR). Yet, these approaches can segue to non-monotonic unexpectedly. Moreover, the much-desired CBR and IndR approaches are at much higher risk for segueing to non-monotonic (as they are already cautiously monotonic and weakly monotonic, respectively). Finally, some of the desired "more sophisticated approaches" (e.g., AnaR, AbdR) squarely reside within the non-monotonic realm, as is evidenced by Table XVII.

Within the realm of analogical reasoning, isomorphism exhibits some promise for being computationally tractable, and certain approaches leverage potentially more accelerated pathways. The amalgam of ICSM2, MAM, and CAIR-A is one such triumvirate approach.

1) ICSM2

As noted in Section IB, the LHM is buttressed by, among other modules, an ICSM2, which determines (should it be needed) antecedent occurrences of IsoPs. In graph theory, if there exists a one-to-one correspondence between the vertice set of S and S', then S and S' are isomorphic. This is shown in Table XVIII, and should the reader desire, this can be affirmed via a variety of tools, such as the one available at https://graphonline.top/en/?graph=xPLjwOkrglDRgYeS. Rather than the graphs, adjacency matrices can also be utilized to determine isomorphism. This can be affirmed via a variety of tools, such as the one available at https://graphonline.top/en/create_graph_by_matrix.

TABLE XVIII. EXEMPLAR ISOMORPHISM BETWEEN S AND S'

Isomorphism between S and S'	***Graph S***	***Graph S'***
f(a) = 1 *f(b)* =2 *f(c)* =3 *f(d)* =4 *f(e)* =5 *f(f)* =6 *f(g)* =7 *f(h)* =8 *f(i)* =9		

Prior work on "graph theory isomorphisms" or Graph Isomorphism (GI) included aberration detection as pertains to the smart grid. Exemplar publications on this thematic are shown in Table XIX.

TABLE XIX. EXEMPLAR DIGITAL OBJECT IDENTIFIERS (DOI) PERTAINING TO GRAPH ISOMORPHISM (GI)

Facet	***DOI***
GI	• 10.1109/IEMCON.2019.8936241 • 10.1109/IAICT62357.2024.10617473 • 10.1109/GEM61861.2024.10585580 • http://dx.doi.org/10.2139/ssrn.5183492 • http://dx.doi.org/10.2139/ssrn.4984663

For that particular Line of Effort (LOE), various aberration detection techniques were utilized, and these are shown in Table XX below.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE XX. EXEMPLAR DOIS PERTAINING TO ABERRATION DETECTION TECHNIQUES

Short-Form Acronym for various Aberration Detection Techniques	***Long-Form Name***
OUV	Over/Under Voltage
OUF	Over/Under Frequency
PJD	Phase Jump Detection
RCAP or dP/dt	Rate of Change of Active Power
ROCOF of df/dt	Rate of Change of Frequency
VVS	Voltage Vector Shift
vTHD	Voltage Imbalance and Total Harmonic Detection (THD)
ROCPAD	Rate of Change of Phase Angle Difference
IM	Intelligent-based Method
EM	Estimation-based Method
WM	Wavelet Transform-based Method

The techniques presented in Table XX were then sorted by Non-Detection Zone (NDZ) and Speed, wherein green indicates either a good NDZ or speed, and red indicates either a poor NDZ or speed, such as shown in Table XXI, row 2. This is then recast in radar chart form, which is shown in Table XXI, row 4. For this particular comparison, the main point to convey, for the presented example in Table XXI, is that the structure itself holds no meaning, as only the values associated with each technique has any significance. Hence, no IsoP testing/comparison needs to be performed (with the associated savings of computational resources).

TABLE XXI. EXAMPLE WHEREIN THE STRUCTURAL TOPOLOGY IS IRRELEVANT; NO ISOP IS NEEDED

Techniques sorted by NDZ

Sorted by NDZ

Technique	NDZ	Speed
OUV	7	5
OUF	7	5
PJD	7	5
RCAP	7	4
ROCOF	7	4
VVS	6	3
vTHD	3	4
ROCPAD	3	2
IM	2	1
EM	2	2
WM	1	1

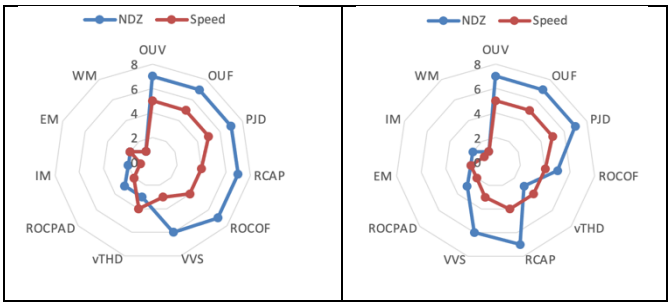
Radar Chart of Techniques characterized by NDZ and Speed (sorted by NDZ)

Techniques sorted by Speed

Sorted by Speed

Technique	NDZ	Speed
OUV	7	5
OUF	7	5
PJD	7	5
ROCOF	5	4
vTHD	3	4
RCAP	7	4
VVS	6	3
ROCPAD	3	2
EM	2	2
IM	2	1
WM	1	1

Radar Chart of Techniques characterized by NDZ and Speed (sorted by speed)



For many ecosystems, wherein addressing computational forays with graph-based approaches can be invaluable, "non-graph-based approaches" can be of comparable value-added proposition (VAP). By way of example, the robust treatment of various figures of speech (e.g., simile, metaphor) and argument (e.g., analogy) (a.k.a., collectively, elements of figurative language) can be invaluable, such as reviewed in Table XXII.

TABLE XXII. EXEMPLAR ELEMENTS OF FIGURATIVE LANGUAGE

Exemplar Element	***Description***
Simile	is a comparison of two disparate entities, via words, such as "like" or "as."
Metaphor	is a direct comparison and asserts that two disparate entities are the same, via words, such as "is," "was," etc. (wherein the words "like" or "as" are not utilized).
Analogy	creates a comparison of how a seemingly disparate entity is akin to, relates to, or is similar to another disparate entity for the purpose of explaining/demonstrating.
Allegory	embodies a more complex/symbolic comparison and leverages a narrative to convey an abstract notion/concept; Wearing points out that "allegories make sense when they're interpreted literally while (most) metaphors do not" [47]. Holme asserts that "arguably, an allegory frames all the events of a story inside an extended metaphor" [48]. More simplistically, Burton claims that "an allegory is a complete narrative that seems to be about one thing but is actually about another" [49].

Hofstadter has argued that analogy is "the core of cognition," and Holyoak seems to affirm [50][51]. Hofstadter further states, "without concepts there can be no thought, and without analogies there can be no concepts" [52]. Given the significance of analogies, it seems prudent to review analogies at various levels of intricacy. Wijesiriwardene asserts that analogies can be viewed "at four distinct levels of complexity: lexical, syntactic, semantic, and pragmatic," and he further notes that "as the analogies become more complex, they require increasingly extensive, diverse knowledge beyond the textual content, unlikely to be found in the lexical co-occurrence statistics that power LLMs" [53].

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

A discussion on Large Concept Models (LCMs), to assist in this regard, can be found within those DOIs shown table XXIII below.

TABLE XXIII. EXEMPLAR DOIS/TITLES OF PAPERS PERTAINING TO LARGE CONCEPT MODELS (LCMS)

Facet	*DOI*
LCM	• 10.1109/IAICT65714.2025.11100570
	• A Prospective Monotonic/Non-Monotonic Transition Zone Impediment for Concept Model-Centric Artificial Intelligence Systems

Wijesiriwardene depicts the increasing levels of analogy complexity, and a version is shown in Figure 3 below.

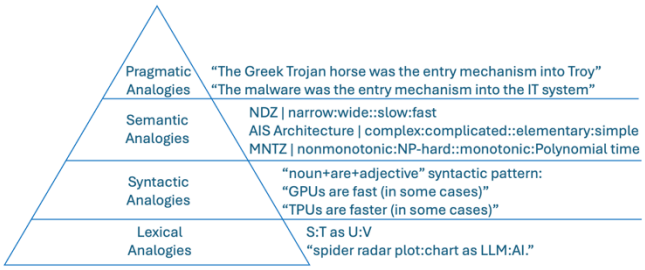


Figure 3. "Increasing Levels of Analogy Complexity" [53].

These increasing levels of analogical complexity are further delineated in Table XXIV.

TABLE XXIV. EXEMPLAR ANALOGY TYPES

Analogy Types	*Descriptors*
Lexical Analogies	Lexical analogies center upon the affinity and shared relationships between concepts despite the concepts being quite distinct and disparate. An example would be S:T as U:V or "spider radar plot is to chart as LLM is to AI."
Syntactic Analogies	Syntactic analogies undertake reasoning by comparison so as to discern structural affinity and relationships between syntactic structures. An example would be that of the "noun+are+adjective" syntactic pattern, which can be seen in the following: "Graphics Processing Units (GPUs) are fast (in some cases)" and "Tensor Processing Units (TPUs) are faster (in some cases)."
Semantic Analogies	Semantic analogies center upon the affinity and shared relationships between entities (while the underlying structure may not be the same). Examples include antonymy (e.g., narrow:wide::slow:fast), synonymy (e.g., complex:complicated::elementary:simple), cause & effect (nonmonotonic:NP-hard::monotonic:Polynomial time), etc.
Pragmatic Analogies	Pragmatic analogies are often predicated upon figurative language, wherein literal translations may not convey the implied meaning intended. A classic example would be, "Reading this paper was a piece of cake;" in essence, the intended meaning was that the paper was easy to read rather than the literal translation that the paper was an edible slice of cake. Wijesiriwardene's example of the Trojan horse theme and comparing it to malware is particularly appropriate for modern times.

Wijesiriwardene explains that "pragmatic analogies are the most complex [of the analogy types], spanning several sentences (often a paragraph) that elaborate on both the source and target domains, contain multiple concept or entities related by diverse relationships, contain abstractions (modeled as subgraphs), and require us to map concepts/entities, relationships and subgraphs between source and target contextualized by external knowledge and a purpose" [53]. To further explain this, if S and S' are isomorphic, then they are analogous in their underlying linkages/relationships and structure. However, while the various forms of analogies exhibit similarities, there may not be a seamless structural match. It then follows that while the gamut of isomorphic cases can be considered analogies, not all analogies are isomorphic. This particular determination is central. If IsoP can be avoided, then the expenditure (of an unknown level of computational resource expenditure) can be bypassed. However, if an IsoP comparison is indeed required, then the ICSM2 operationalizes the IsoP comparison via the derivation of a specific amalgam, which is comprised of two key facets, such as shown in Figure 4 below.

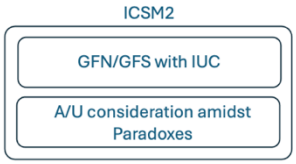


Figure 4. ICSM2's Derivation of a specific amalgam: (1) a pertinent GFN/GFS and Inherent Uncertainty Construct (IUC), and (2) A/U Consideration Amidst Paradoxes

Figure 4 needs some unpacking. First, starting with the more discrete facet, Generalized Fuzzy Numbers (FN) (GFN) can be quite useful. The type of shape of the involved GFN is typically captured by the "*f-sided*" notation within the long form of "Generalized *f-sided* Fuzzy Number" as well as by "*f*" notation in the short-form of the "G*f*FN" acronym. The G*f*FN membership function can extend into a more sophisticated shape beyond the more common and simplistic 3-sided (i.e., triangular) or 4-sided (i.e., trapezoidal, which has a minimum of one pair of parallel sides) prototypical shapes. Examples of the involved notation for these more sophisticated shapes are shown in Table XXV below.

TABLE XXV. EXEMPLAR GENERALIZED FUZZY NUMBERS (GFNS) WITH SHAPE TYPES

Short-Form Acronym	*Long-Form Name*
G***Tr***FN	Generalized ***Triangular*** Fuzzy Number

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

G***Tp***FN	Generalized ***Trapezoidal*** Fuzzy Number
G***Pe***FN	Generalized ***Pentagonal*** Fuzzy Number
G***Hx***FN	Generalized ***Hexagonal*** Fuzzy Number
G***Hp***FN	Generalized ***Heptagonal*** Fuzzy Number
G***Oc***FN	Generalized ***Octagonal*** Fuzzy Number

In some cases, within the literature, the *Tr, Tp, Pe, Hx, Hp, Oc*, etc. are supplanted by "*N*," and the utilized acronym is that of G*N*FN. Regardless, the *"N-sided"/"f-sided"* nature allows the membership function to be defined by a greater multiple of linear and/or non-linear functions for better capturing/representing the U of A/U than that by the simpler [prototypical] shapes (given their limitations with regard to their constituent number of linear and/or non-linear functions able to be represented). However, a further distinction is also made. A G*f*FN can not only have a varying number of *f* sides, but also the sides comprising the boundary of the G*f*FN might not necessarily be linear (i.e., the sides can be non-linear, such as in the case of a Gaussian curve, etc.). Accordingly, the notation can be of the form GL*f*FN for a Generalized Linear *f-sided* FN and GNL*f*FN for a Generalized Non-Linear *f-sided* FN. Taking the example case of 5-sides, GL*Pe*FN denotes the involved "Generalized Linear *Pentagonal* FN," wherein the sides are all linear; in contrast, GNL*Pe*FN denotes Generalized Non-Linear *Pentagonal* FN, wherein the side(s) might be non-linear. Chakraborty depicts a GL*Pe*FN in [54], and a variation is shown in Figure 5.

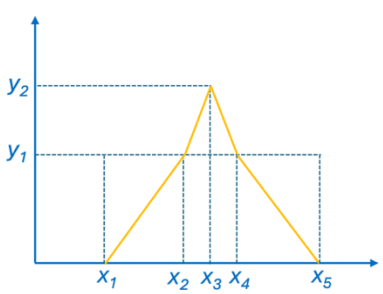


Figure 5. An exemplar GLPeFN

Chakraborty also provides an example of a GNL*Pe*FN, and a version is reproduced below in Figure 6 [54].

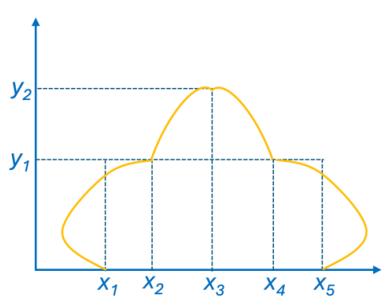


Figure 6. An exemplar GNL*Pe*FN

In some cases, the GL*Pe*FN approximation can suffice, rather than computing the GNL*Pe*FN, and Velu & Ramalingam provide an example of this, for which a variation is shown below in Figure 7 [55].

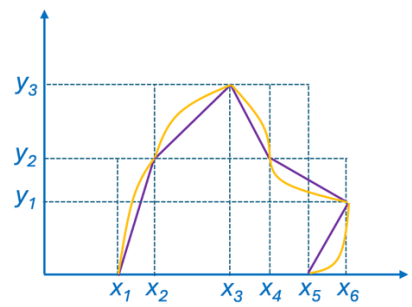


Figure 7. An exemplar GNL*Pe*FN

Before plowing forward, some background information is necessary; an Interval-Valued Fuzzy Number (IVFN) is a fuzzy number, wherein the degree of membership is denoted by an interval (as a range of prospective membership values) instead of a single value. Along this vein, Chakraborty further presents a Generalized Linear Interval-Valued *Pentagonal* Fuzzy Number (GLIV*Pe*FN), and a variation is shown in Figure 8 (with symmetry) along with a *Hexagonal* rendition (GNLIV*Hx*FN) in Figure 9 (with asymmetry) [54].

Figure 8. An exemplar GLIV*Pe*FN (with symmetry)

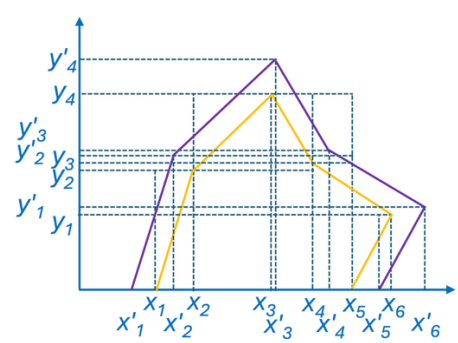


Figure 9. An exemplar GLIV*Hx*FN (with asymmetry)

Chakraborty further presents a Generalized Non-Linear Interval-Valued *Pentagonal* Fuzzy Number (GNLIV*Pe*FN),

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

but to be consistent with Figure 9, a Hexagonal rendition (GNLIV*Hx*FN) is cast below in Figure 10 (with asymmetry).

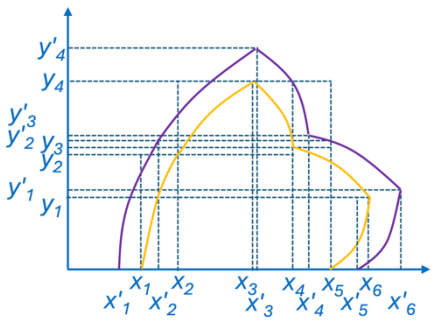


Figure 10. An exemplar GNLIV*Hx*FN (with asymmetry)

Again, as previously shown in Figure 7, for some cases, best-fit approximations (or even more rudimentary approximations) may suffice. This raises the issue of which particular approach vector might be prudent (depending upon the need). Prior to unpacking this, Intuitionistic Fuzzy Numbers (IFNs) will be explained. In essence, IFN embody both membership and non-membership degrees, thereby providing a more nuanced delineation of uncertainty. Mert depicts a Generalized Non-Linear P [Intuitionistic Logic] *Pentagonal* Fuzzy Number (with asymmetry), but to be consistent with Figures 9 and 10, a hexagonal version is cast in Figure 11 [55].

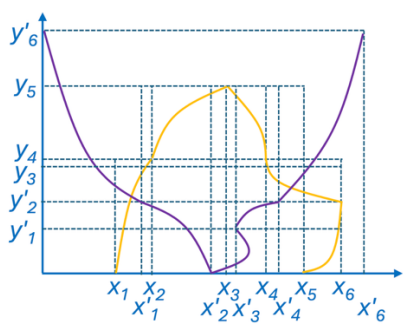


Figure 11. An exemplar GNLIL*Pe*FN (with asymmetry)

As can be gleaned, IFNs and IVFNs depict U in varying ways. For some cases, IFNs are able to be expressed in IVFN form (e.g., when membership as well as non-membership values are expressed as single-point intervals), such as in the case wherein (0.9, 0.1) segues to ([0.9, 0.9], [0.1, 0.1]). In contrast, the reverse is not necessarily true, such as in the case, wherein ([0.1, 0.9], [0.3, 0.7]) might not necessarily translate to a "single, representative IFN." For certain basic operations (e.g., "and", "or"), equivalence can be demonstrated. However, this is not necessarily the case for more complicated operations.

Proceeding from the various GFNs to the broader category of Generalized Fuzzy Sets (GFSs) can be quite useful as well. By way of context, the classical FS involves a membership function that assigns a degree of membership (e.g., between 0 and 1) to each element within the involved FS. However, this does not necessarily reflect the *hesitation* for scenarios with incomplete/uncertain information. In contrast, GFS encompass a range of extensions to the classical Fuzzy Set (FS), such as that of Intuitionistic FS (IFS), Interval-Valued FS (IVFS), etc. Along this vein, IFR equates to Intuitionistic Fuzzy Reasoning and IVFR equates to Interval-Valued Fuzzy Reasoning. In terms of isomorphism, IFSs and IVFSs have been shown to be formally equivalent [56][57]. In essence, they share equivalent mathematical structures in spite of their differences in representing U. Luo had expressed this diagrammatically, for which a version is reproduced in Figure 12 [57]. With regards to Figure 12, TIA equates to the Triple I Algorithm (i.e., "a fuzzy reasoning algorithm," which posits an output given a specified input), RTIA equates to the Reverse Triple I (a.k.a., R-III) Algorithm (i.e., "a fuzzy reasoning algorithm," which inverts the TIA so as to ascertain an apropos input given a specified desired output), SIS equates to Subsethood Inference Subsethood (i.e., the degree of containment of a FS within another FS), *b* equates to bijection, and *i* equates to isomorphic [57].

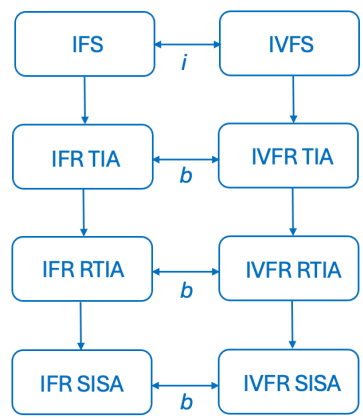


Figure 12. The IFS and IVFS Isomorphism and various Bijections [57]

The significance of this resides in the paradigm that, generally, IFS tend to be quicker to compute than IVFS (particularly in the area of basic comparisons/operations). Moving beyond IVFS, the Interval-Valued Intuitionistic Fuzzy Set (IVIFS) yields even "more precise results" [58]. After all, while an IVFS depicts the uncertainty of membership via an interval, an IVIFS leverages both an interval-centric degree of membership as well as an interval-centric degree of non-membership. Thus, IVIFS is even more nuanced than IVFS. Fortunately, it is a generalization of IFS, which is isomorphic to IVFS, as Bustince well depicted, and a version is reproduced in Figure 13 [118]. Along the vein of Figure 13, the Type 2 Fuzzy Set (T2FS) had been discussed in, among other papers, the second DOI of Table IV, which discussed the notion that the "Type-2 Fuzzy Set (T2FS)…can accommodate the uncertainty of membership fluidity, whereas the Type-1 Fuzzy Set (T1FS) only accommodates membership invariableness" [second DOI of Table IV][first

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

DOI of Table VIII][second DOI of Table VIII]. Along with IVFS, IFS, IVIFS, there are also the notions of Vague Set (VS) and Grey Set (GS).

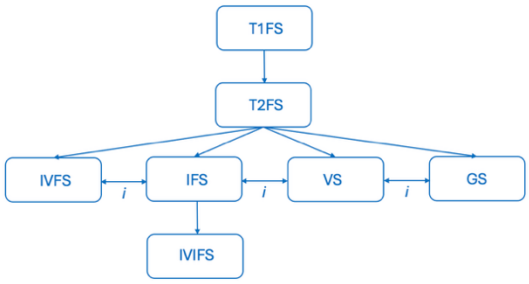


Figure 13. Mathematical Equivalence among particular GFS [118]

Lu illuminates the notions that VS "is more natural than an IFS for merging fuzzy" values [59]. A classic example involves the merging of three fuzzy values, such as 0.3/*m*, 0.7/*m*, and 0.9/*m*; the resultant VS expression would be [0.3, 0.9]/*m*, wherein the lower bound of the membership *m* is 0.3 and the upper bound is 0.9. In contrast, the IFS expression would be less intuitive with 0.6 for the degree of membership, 0.2 for the degree of non-membership, and 0.2 for the hesitation margin. Furthermore, Lu notes that ascertaining the Median Membership (M_m) and/or the Imprecision Membership (M_i) is more intuitively gleaned via VS, such as depicted in Figure 14, which is a variation of Lu's figure.

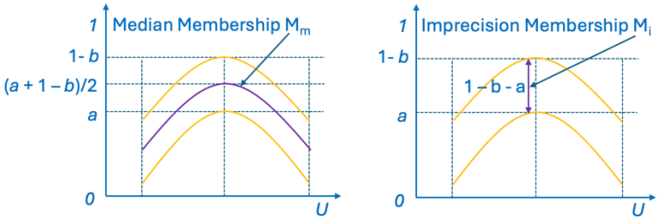


Figure 14. Delineating the Median and Imprecision Membership via VS, where y axis denotes "Membership Functions" and the x axis denotes "Data Objects" [59]

Likewise, delineating the hesitation region for VS is much more straightforward, such as shown in Figure 15. In essence, VS is more intuitive for delineating the Support Region, Opposition Region, and the Hesitation Region (wherein there is neither support nor opposition).

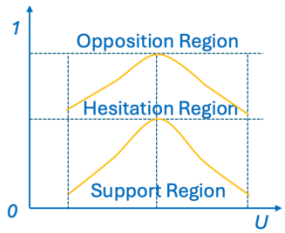


Figure 15. Delineating the Hesitation Region for VS (as contrasted to IFS)

As Lu discusses, "in the literature, the notions of IFSs and VSs are regarded as equivalent," in the sense that an IFS is isomorphic to a VS [59]. This is useful for transitioning among IVFS, IFS, VS, and GS [118]. Next, there is also a distinction between VS and GS. For example, Alkhazaleh states that a VS "is defined by a truth-membership function … and a false-membership function" [60]. Lu adds to this by asserting that "interval-based membership is used in a VS" and that "the interval-based membership generalization" in VS "is more expressive in capturing vagueness of data" [59]. On the other hand, GS leverages the notion of Grey Numbers (GNs), and Khuman notes that "the generalised" GN "can cater to both discrete and continuous data" [61]. Although there is the obvious connection between GS and GN, they are distinguished by the fact that a GS captures uncertainty about membership to the involved set while a GN denotes uncertainty about the actual value of a particular quantity. Sifeng provides a generalization for both: "the information that grey is often associated with is information that can be described as being partially known and partially unknown, which in actuality is a common occurrence of uncertain systems" [62]. Along this vein, Shaker and Moore-Clingenpeel reference the epistemological (pertaining to the theory of knowledge) constructs of the "known knowns" (KK), "known unknowns" (KU), and "unknown unknowns" (UU) that were infamously popularized by Donald Rumsfeld. The extrapolated quad chart, which also references the writings of Shaker and McGregor, is shown in Table XXVI [63][64].

TABLE XXVI. EPISTEMOLOGICAL CONSTRUCTS [63][64]

Known Knowns (KK) "Things we are aware of and understand"	*Known Unknowns (KU)* "Things we are aware of and do not understand"
Unknown Knowns (UK) "Things we are not aware of, but understand"	*Unknown Unknowns (UU)* "Things we are not aware of and do not understand"

With regards to Sifeng's referencing of "partially known" and "partially unknown," this could pertain to the KK and KU for the former and KU, UK, and UU for the latter. As previously stated in the prologue of Section II (and as exhibited in Tables XI, XII, XVI, XVII and the broader categorizations of Table XIV), despite the varied RMs that might be utilized, an inconsistency/contradiction could occur due to phenomenon, such as the Ellsberg Paradox, which can be described as "ambiguity aversion" — a predilection to avert alternatives whose prospects are unknown; citing the logic of Lang ("zero exposure to ambiguity can be optimal") as well as the assertions of Ellsberg, Machina, and others, ambiguity aversion "violate[s] both the key rationality axioms and classic models of choice under uncertainty" [65]. For a number of cases, ambiguity aversion segues to sub-optimal DM [65][66]. As a case in point, Jia's findings show that

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

experimental participants preferred KK over KU/UK/UU even when the options with U might be more favorable (it is not clear whether KK/KU prevailed over UK/UU). In any case, Jia does indeed note that "participants who learned about the Ellsberg Paradox were more tolerant of ambiguity, yet ambiguity aversion was not completely abolished" [67]. Coleman summarizes matters as: "The Ellsberg paradox is often cited as evidence for unknowable 'ambiguity' versus computable 'risk' and a refutation of the Savage axioms regarding expected utility maximization" [68][69]. Chen describes the Ellsberg paradox more simply, as "people prefer betting on known (objective) probabilities rather than unknown (subjective) probabilities" [70]. Weber describes it as "in *ambiguity* over time, the eventual outcome is known, but the length of time before the outcome will occur is *uncertain* [71]. In Table I of Section I, it is clear that ambiguity and uncertainty are indeed different. Yet, despite these differences, definitions of the "Ellsberg paradox" demonstrate how these terms are often used interchangeably (with a lack of distinction made between the two). Exemplars of this are shown in Table XXVII.

TABLE XXVII. EXEMPLAR UTILIZATION OF AMBIGUITY (A) VERSUS UNCERTAINTY (U) IN VARIED DEFINITIONS OF THE ELLSBERG PARADOX

Researcher	***Remarks***
Binmore	"Experimental results on the Ellsberg paradox typically reveal behavior that is commonly interpreted as *ambiguity* aversion" [72].
Leopold	"Smarter in the Long-Term: Diminishing *Ambiguity* Aversion in a Repeated Ellsberg Urn Task" [73].
Chen	*"Ambiguity* Aversion: The Ellsberg paradox shows that people prefer betting on known (objective) probabilities rather than unknown (subjective) probabilities" [74].
Halevy and Feltkamp	"The Ellsberg paradox demonstrates that people's belief over *uncertain* events might not be representable by subjective probability" [75].
Segal	"Measuring Nonmonetary Utilities in *Uncertain* Choices: The Ellsberg Urn" [76].
Joreno-Jimenez and Vargas	"we present a method to deal with *uncertainty*, which considers Ellsberg's objections" [77].
Jabarian and Lazrus	"55% of the subjects prefer avoiding *ambiguity* even when it means choosing dominated risky options – what we call the Two-Ball Ellsberg Paradox" [78].

Perhaps, the preferred embodiment for the use of the cited terms (ambiguity, uncertainty), with regards to the Ellsberg paradox, is as follows. In essence, the Ellsberg paradox alludes to the notion that people have a predilection towards *choices with less uncertainty*, thereby running counter to the Expected Utility Theory (EUT) (e.g., a rational cost-benefit analysis) by subscribing to *ambiguity aversion*.

Apart from the Ellsberg paradox, there is also the Machina paradox (wherein the preference ranking may change although the underlying probabilities are equivalent). Aerts notes that the Machina and Ellsberg paradoxes run counter to the EUT [79]. Likewise, the Allais paradox (where DM are unlikely to consistently make rational decisions under CDC) runs counter to the EUT [80]. Both the Machina and Allais paradoxes run counter to Savage's Independence Axiom (IA). Ferrari-Toniolo proclaims that the IA is the "most demanding axiom" as pertains to EUT [81]; Blavatskyy explains IA simply: "The independence axiom postulates that"…the DM's… "preferences between two lotteries are not affected by mixing both lotteries with the same third lottery (in identical proportions)" with the extension of this being that the DM "does not necessarily prefer the same choice alternative when repeatedly presented with the same choice set" [82]. Overall, the Ellsberg, Machina, and Allais paradoxes all run counter to EUT and highlight DM behavior that deviates from rational behavior, such as is shown in Table XXVIII.

TABLE XXVIII. EXEMPLAR TYPES OF PARADOXES WITH TYPE OF EFFECTS

Type of Paradox	***Type of Effect***	***Description***
Ellsberg	Ambiguity Aversion	DMs have a predilection to be ambiguity-averse and tend towards choices with known calculable risks rather than those with unknown incalculable risks.
Machina	Inconsistent Preference Rankings	Accurate preference ranking methods are problematic amidst A and U; this has high impact for AI models that are trained to learn and predict preferences.*
Allais	Certainty Effect	DMs have a predilection for more certain outcomes over probabilistic outcomes.

Jim refers to this as "each of these models is trained from a common base model to predict the…preferences of a single individual," Aldoseri refers to this as "machine learning algorithms learn, make predictions, and improve their performance over time" [83][84]. From Table XXVIII and the discussions leading up to this point, the key take-away, taking the cue on inconsistent preference rankings, is: (1) with regards to non-monotonic phenomenon, human behavior may not be able to be captured and the involved AIS may actually adopt "the opposite" position (from Section IIB) [34]; (2) "the shift of the involved variables from a monotonic to a non-monotonic paradigm can be quite unexpected and occur more frequently than anticipated/desired" for AIS (from Section IIC) [2]; and (3) as report by Chen, with regards to inconsistent preference rankings, "most state-of-the-art preference-tuned models achieve a ranking accuracy of less than 60% on common preference datasets," which basically equates to "preference

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

learning algorithms do *not* learn preference rankings," "existing reference models *rarely have* correct rankings," and "preference learning *rarely corrects* incorrect rankings" [85]. Also, interestingly and ironically, leaning the other way towards monotonic does not seem to be viable either; for example, "enforcing a strict monotonic paradigm can segue to an unnatural rigidity and/or incorrect/irrelevant responses by the" AIS (from Section IIC). A robust counterpoising seems to be the key.

III. EXPERIMENTATION

Counterpoising mechanisms, such as that of a robust MMSO mechanism, have been previously experimented upon, as shown in Table XXIX.

TABLE XXIX. EXEMPLAR DOIS AND PAPER TITLES PERTAINING TO THE NOTIONS OF MADM/MODM SM/OM

Facet	*DOI*
MADM/MODM SM/OM (MMSO)	• A Prospective Monotonic/Non-Monotonic Transition Zone Impediment for Concept Model-Centric Artificial Intelligence Systems • 10.1109/IAICT65714.2025.11100570 •10.1109/AIIoT65859.2025.1110531 • http://dx.doi.org/10.2139/ssrn.5291883 • http://dx.doi.org/10.2139/ssrn.5291881 • http://dx.doi.org/10.2139/ssrn.5291879 • 10.1109/ICAIIC64266.2025.10920828 • http://dx.doi.org/10.2139/ssrn.5183492 • 10.1109/IAICT62357.2024.10617473 • 10.1109/GEM61861.2024.10585580 • 10.1109/AIIoT61789.2024.10579033 • http://dx.doi.org/10.2139/ssrn.4984663 • 10.1109/ICSGTEIS60500.2023.10424230 • 10.1109/ICPEA56918.2023.10093212

As previously discussed in Section IID, "depending upon the RM/amalgam of RMs/overall RM pathway initially selected, such as by the LHM, the related DM point/DMIP will likely impact the downstream behavior within the MNTZ in terms of a preponderance of monotonic or non-monotonic behavior." Consequently, the selection is non-trivial. To assist in this regard, the MMSO could be quite useful. After all, as pertains to preference ranking, it is a central component of MMSO, which is the mainstay of MCDM. For MCDM, preference ranking impacts the specification of alternatives, the defining of criteria, the assigning of weights, and the sorting of alternatives, among other items.

Homing in on the aforementioned delineation of criteria, assignment of weights, and the ranking of alternatives, the Shapley value has been found to be quite instrumental. It can facilitate the analysis of interactions between the various criteria, thereby providing invaluable insights during the ascertainment of apropos criteria weights, as well as enhance various ranking methods. As Qin puts it, the Shapley value "is widely used for" "feature importance analysis" [86]. To further this, B. Rozemberczki asserts that, "the Shapley value is used to measure the contributions of input features to the output of a machine learning model" [87]. In essence, the Shapley value can be construed as helping to gauge the overall significance of each criterion. Given this pivotal role, it seemed prudent to explore *the Shapley value in terms of its prospective value within the MNTZ for the ultimate purpose of helping to better counterpoise MR/NMR via a better counterpoised MMSO*. For this reason, explorations of a bespoke IbSOA approach were undertaken.

Given the various counterpoising within the discussed AIS, enhanced context is a mainstay of this paper, and this was alluded to in [1], which this paper is rooted upon. Axiomatically, to enhance context for an AIS, ML2 is vital. Hence, in the case of this paper, interstitial analyses are crucial for ascertaining "whether the prospective ML learnings are of potential benefit" [1]. In terms of theoretical foundations, it is posited that "Borgonovo's glocal notion can help bridge the gap, and the significance of the [Optimal Shapley-Nondominated Solution] OSNS segueing to an Optimal Shapley-Owen-Nondominated Solution (OSONS)" paradigm is underscored in [1] as well as articulated by various researchers within this arena, such as Casajus, Lopez, Beal, and others [88][89][90]. Fundamentally, the referenced Owen value nicely captures the intricate interactions between/among the constituent members of the involved feature set as well as "extends the Shapley value (which well captures the individual feature contributions) in a consistent fashion" [1]. Yet, OSONS is predicated upon the referenced IbSOA, which encompasses "temporal-centric [Finite-Change Shapley-Owen] FCSO values, ["Squared Cohorts" Shapley-Owen] SCSO values, and [Generic Shapley-Owen] GSO values/[Variance-Based Shapley-Owen] VBSP values" [1]. In addition, to operationalize the IbSOA, it should be conjoined with the previously referenced MMSO mechanism.

Based upon the experimentation referenced in Table XXIX, it seemed prudent to further investigate those areas promulgated by Wu, Wang, Hua, and others. By way of example, among others, Wu and Wang have been proponents of "Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) OM in conjunction with SHAP," and this seemed favorable for MMSO experimentation [91][92]. As another example, among others, Hua has been a proponent of "the PROMETHEE OM with SHAP" [1]. Again, this seemed promising for MMSO experimentation as well. As noted in [1], the experimentation assessed the referenced favorable/promising avenues, and an output, among others, centers upon the notion that it seems well advised to utilize "an OM (e.g., CRiteria Importance through Intercriteria Correlation or CRITIC) to first, derive the criteria weights and second, use a complementary pairing for the ensuing ranking (e.g., TOPSIS, PROMETHEE)" [1]. The reader

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

should note that a substantive portion of the balance of Section III (particularly Sections IIIA and IIIB) is derived from [1], and that this journal paper is an invited extended version of that paper (i.e., [1]).

A. Experimental Testbed

For a tasked ML to well learn atop another ML in an ML2 sense, it is important to mitigate against inadvertent spawning (i.e., *Spawn Reduction*). Accordingly, enhanced context and understanding at the interstices can be central to the necessary ML2 mechanism to substantially decrease the "spawning of further non-convex MINLP (e.g., from the transformation pathways of non-convex MINLP to convex MILP)" [1]. For the experimentation in [1] and herein, "SDP solvers were implemented aboard GNU's Not Unix (GNU) Octave (a "numerical computation platform" that is "under the GNU [General Public License] (GPL) v3 license" and is generally "compatible with the likes of MATLAB") along with a myriad of Octave Forge packages" [1][93]. As noted in [1] and [93], "'the source code was modified in the lab environment' so as to implement accelerants for the referenced SDP solvers to quickly address the various involved convex optimization problems described herein;" also, as noted in [1] and [93], "GPLv3 avoids the issue of tivoization (the instantiation of a system that incorporates software under the terms of a copyleft software license but leverages hardware restrictions or digital rights management to prevent users from running modified versions of the software on the involved hardware)" [93]. The experimental "testing was conducted using a variety of open-source software packages, such as Automatic Differentiation Model Builder (ADMB) (for non-linear statistical modeling) and Interior Point OPTimizer (IPOPT) (for large-scale nonlinear optimization)" [1][93]. Additional "promising software packages, such as LOQO (like IPOPT, it is based upon the interior-point method) and Sparse Nonlinear OPTimizer (SNOPT) (it leverages Sequential Quadratic Programming or SQP for resolving large-scale non-linear optimization problems) were examined, but they were not utilized given their licensing caveats" [1].

It had been discussed in [1] and [the second DOI of Table IV] that a particular numerical implementation of Continuous Wavelet Transforms (CWTs), aboard a **C**onstriction Factor (CF)-**P**article Swarm Optimization (PSO)-**R**obust Convex Relaxation (RCR)-**L**ong Short-Term Memory (LSTM)-**D**eep Convolutional Neural Network DCNN (CPRLD) architectural paradigm, well contributes to System Transparency Explainability & Accountability (STEA) by way of the intrinsic "successive convolutional layers (which contain the cascading of ever smaller 'CWT-like' convolutional filters)" [1][the second DOI of Table IV]. As noted in [1], "the referenced CPRLD construct handled the various transformation pathways" alluded to in Figure 2 previously "(e.g., convex approximations, series of convex relaxations, etc.), and the architectural implementation for this paper was unique in that a ML2 paradigm was implemented for *Spawn Reduction* (SR_2 on SR_1)," such as depicted in Figure 16.

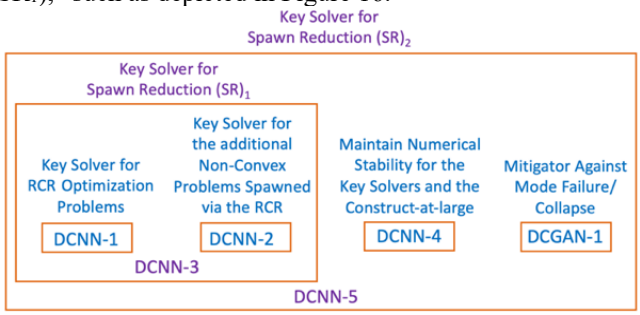


Figure 16. CPRLD Architectural Construct with a ML2 (SR_2 on SR_1) Spawn Reduction paradigm [1]

As discussed in [1], "in terms of implementation details, a DCNN-centric instantiation was chosen for the requisite sufficient balance of reduced computational complexity along with sufficient robustness to be fit for purpose. The assigned tasks of the various DCNN are labeled accordingly in Figure 16. For example, as DCNN-1 was tasked with being the key solver for the involved convex optimization problems, it required a high degree of numerical stability, and PyTorch version 0.4.1 was selected; DCGAN-1 leveraged a "forward stable" TensorFlow-based Deep Learning (DL) Convolutional GAN (DCGAN) implementation to be able to well address the potentiality of mode collapse/mode failure (a phenomenon that may occur when adversarial GANs, which are being trained in tandem, are either unable to converge or undergo an anomalous convergence)" [the second DOI of Table IV].

B. Experimental Construct

As pertains to the involved experimental construct, which is based upon Figure 4 of [1], such as is now shown in Figure 17, "prior experimentation aspects used as presets are reflected in blue font while current experimental elements are shown in purple font" [1].

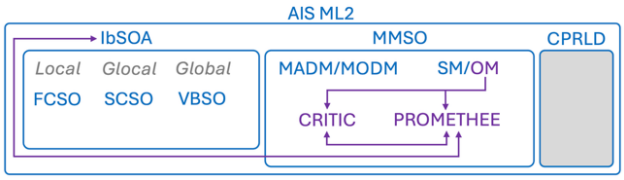


Figure 17. Conjoining of IbSOA and MMSO

For additional context, "t-" elements (e.g., t-FCSO, t-SCSO, t-GSO, t-VBSO) of the b-SHAP can be extrapolated, and these relate to Borgonovo's work…which more fully considers "Kotthoff's emphasis on temporal-sensitive/temporal-centric Shapley values" [94]. As described in [1], "the OM of CRITIC was utilized as a preset for deriving the criteria weights, and the OMs of PROMETHEE, TOPSIS, and ÉLimination Et Choix Traduisant la REalité (ELECTRE) were utilized for the

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

subsequent rankings. Initial selections and avoidances, among others, were based upon the following rationale. For example, PROMETHEE was known to be 'easily… understood' and interpretable, so it was selected for testing" [95][96]. Along this vein, [fuzzy] VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) was not selected, as it was known to be less interpretable and "less explainable than other more intuitive methods" [97].

In general, selections were made in adherence with the principles for high efficacy ML2. In addition, there were other technical considerations. As discussed in [1] as well as those papers pertaining to related experimentation, such as shown in Table XXX, the issue of "Rank Reversal" (RR), wherein ranking results might change when the method changes or when the set of alternatives changes (leading to inconsistent and/or inaccurate results), was also investigated (as RR affects several of the methodologies)

TABLE XXX. EXEMPLAR DOIS PERTAINING TO THE NOTIONS OF RANK REVERSAL (RR)

Facet	*DOI*
RR	• 10.1109/IAICT65714.2025.11100570 • 10.1109/AIIoT65859.2025.11105315 • http://dx.doi.org/10.2139/ssrn.5291883 • http://dx.doi.org/10.2139/ssrn.5291881

As noted in [1], "the select OMs experimented with were known to be the most resistant to RR (yet are still subject to the phenomenon), and preliminary results" are reflected in Figure 18 [98]. "The key for the chart is as follows. First, the referenced 'select OMs' of this Section IIIC are self-evident: ELECTRE, TOPSIS, and PROMETHEE" [1] "Second, these 'select OMs' were benchmarked by execution time (E), sensitivity (S), performance under uncertainty (U), validity (V), and interpretability (I)" [1]. "Third, the aforementioned were benchmarked against classical SHAP (c-SHAP), as well as the b-SHAP approach described within this paper" [1]. "Using the CPRLD as a preset, collectively, this forms the basis" of the mechanisms described herein. The relative values were normalized against a scale of one to ten for ease of comparison" [1].

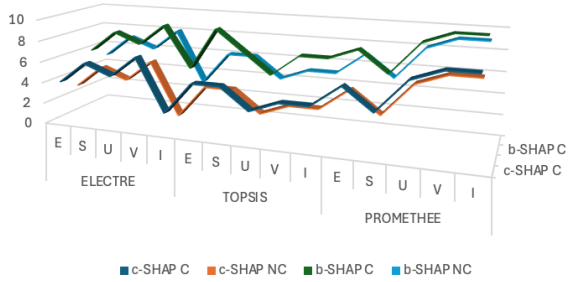


Figure 18. Preliminary Results from b-SHAP/select OM Benchmarking

Also, "the V and I were higher for PROMETHEE than for TOPSIS or ELECTRE. The E for TOPSIS was notably higher than that of the others, but the computational complexity is known to be less, and the performance under conditions of U was weaker than that of the others; the performance of PROMETHEE under conditions of U were seemingly better than ELECTRE and TOPSIS, in that order. Overall, the performance of b-SHAP was better than that of c-SHAP across the board for the range of E, S, U, V, I (for all the "select OMs" of ELECTRE, TOPSIS, and PROMETHEE). Hence, the b-SHAP-PROMETHEE amalgam (along with the CRITIC, CPRLD, etc. presets) exhibits promise" [1].

C. *Discussion re: IbSOA*

As noted in the first DOI of Table XXIX, "the myriads of interplays among local, glocal, and global is clear, as a transformation and/or sequence of transformations can lead from one to another" [94]. For example, what Mase deemed to be the [local] *Baseline Shapley* (i.e., what equates to the average of the *FCSO values* function under uncertainty) can be readily transformed to the [global] *VBSO values* [94][99]. Likewise, the [glocal] *SCSO values* can be transformed to the *VBSO values*. These interplays, among others, demonstrate how "additional insights into the [ML] model behavior" are possible [99].

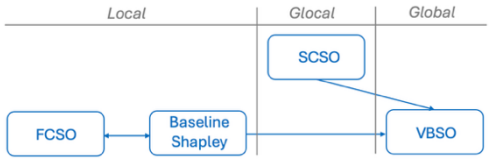


Figure 19. Derived from Figure 1 of [94]; the interplay among Local (FCSO and Baseline Shapley), Glocal (SCSO), and Global (VBSO)

This is akin to the Figure 13 interplays, and the successful operationalization of Figure 17 is contingent upon a phenomenon, which has been previously investigated, such as exhibited in Table XXXI — Robust Convex Relaxations (RCR).

TABLE XXXI. EXEMPLAR DOIS PERTAINING TO THE NOTIONS OF ROBUST CONVEX RELAXATIONS (RCR)

Facet	*DOI*
RCR	• A Prospective Monotonic/Non-Monotonic Transition Zone Impediment for Concept Model-Centric Artificial Intelligence Systems • 10.1109/IAICT65714.2025.11100570 • 10.1109/AIIoT65859.2025.11105315 • http://dx.doi.org/10.2139/ssrn.5291883 • http://dx.doi.org/10.2139/ssrn.5291881 • http://dx.doi.org/10.2139/ssrn.5291879 • http://dx.doi.org/10.2139/ssrn.5183492 • 10.1109/IBDAP62940.2024.10689701 • 10.1109/IAICT62357.2024.10617473 • 10.1109/GEM61861.2024.10585580 • 10.1109/AIIoT61789.2024.10579033 • http://dx.doi.org/10.2139/ssrn.4984663 • http://dx.doi.org/10.2139/ssrn.4679251

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

	• 10.1109/ICSGTEIS60500.2023.10424230 • http://dx.doi.org/10.2139/ssrn.4679260 • 10.1109/ICPEA56918.2023.10093212 • 10.1109/CyMaEn57228.2023.10051057 • 10.1109/CyMaEn57228.2023.10050946 • 10.1109/OETIC57156.2022.10176215 • http://dx.doi.org/10.2139/ssrn.4287248 • https://ssrn.com/abstract=4219298

By leveraging the described RCR approach, the notion of relaxations/convex relaxations can be applied so as to attain "Relaxed Isomorphisms"/"Isomorphic Relaxations." In this way, should IsoP be required, the DMP can be more computationally tractable. Mancinska references this approach as do Atserias and Maneva. Aflalo asserts "that for friendly graphs, the convex relaxation is guaranteed to find the exact isomorphism or certify its inexistence" [100][101][102].

The starting impetus of this paper was to address SDM, particularly SDM-A/U, and the LHM was discussed as a meaningful option. As discussed, the main components of the LHM include the ICSM2 and MAM. The ICSM2 has a list of questions to address (e.g., whether the involved set is UnS, EqS, EquivS, OrS, POSET, UnS-Iso, POSET-Iso, or IV), such as previously described in Table V. The MAM also has a list of questions to answer (e.g., how much time is available, what and how much information is available, as this might dictate that type of initial RM is most feasible and what is the probability that the involved is PDMP, ADMP, or NPDMP), such as shown in Figure 20, and re-expressed in Figure 21 in terms of N+V or IsoP pathways.

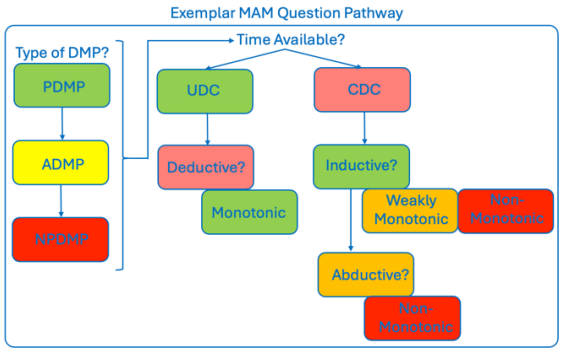


Figure 20. MAM Exemplar Question Pathway

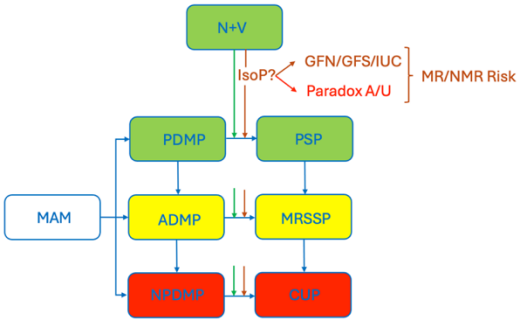


Figure 21. N+V or IsoP Pathways; IsoP has MR/NMR Risk

It should be remembered that Deductive is achieved iteratively (not front-loaded) while Inductive and Abductive are front-loaded. In addition, the CAIR-A also has the challenge of ascertaining what primary RMs and secondary RMs might be involved, and this was previously presented in Table XVII.

As noted in the 1st DOI of Table VIII, "the establishment of an Inherent Uncertainty Construct (IUC)…is central, and crucial to this is the leveraging of T2FS and Zadeh's Fuzzy Systems Theory (a.k.a., IUC-1a) along with the consideration of Rough-Fuzzy Set (a.k.a., IUC-2a), which is an extension of IUC-1a and Pawlak's Rough Set (a.k.a., IUC-1b); after all, IUC-2a can well accommodate the notion of an affiliation, "but not necessarily absolute inclusion" [103]. Furthermore, "Deng's Grey Systems Theory (a.k.a., IUC-2b) can enhance the precision of IUC-2a."

TABLE XXXII. COMPONENTS OF THE INHERENT UNCERTAINTY CONSTRUCT (IUC)

Short-Form Acronym	***Long-Form Name***
IUC-1a	Leverages T2FS and Zadeh's Fuzzy Systems Theory (FST).
IUC-1b	Pawlak's Rough Set (RS).
IUC-2a	Rough-Fuzzy Set (RFS) can well accommodate the notion of an affiliation, "but not necessarily absolute inclusion" [103]. It is an extension of IUC-1a and IUC-1b.
IUC-2b	Deng's Grey Systems Theory (GST) can contend with systems with incomplete or uncertain information. For RWS, the uncertainty can stem from continuous data that may absent, exhibit intricate relationships, be noisy, etc. GST leverages GSs (which rely upon GNs), which are numbers with known upper and lower boundaries, but for which the precise value is unknown within those boundaries; this embodiment is able to concurrently represent of both discrete and continuous data simultaneously) to enhance the precision of IUC-2a. GST also leverages discontinuous Grey Sets, which are specifically geared to handling discontinuous data (wherein only partial information is available) or data with uncertainty. It should be remembered that although a discontinuous function may have a continuous domain, it may have breaks/gaps/jumps or points where it is not defined. In contrast, a discrete function has distinct and disparate values (e.g., integers).
IUC-3	Information Entropy Methods, whose strength resides in ascertaining "unknown attribute weights" [1st DOI of Table VIII][104].

"If the relationship/membership (e.g., entity, attribute, etc.) is discontinuous, IUC-2b can be leveraged; otherwise, given a continuous/continuous alternative paradigm, then other Probability [& statistics] Systems Theory (PST) approaches might be utilized, such as Information Entropy Methods (IEMs) (a.k.a., IUC-3), whose strength resides in

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

ascertaining 'unknown attribute weights'" [second DOI of Table IV][1st DOI of Table VIII][104]. Furthermore, "whether the relationship/membership is discontinuous or continuous (e.g., pulsed, rather than continuous), it can still be construed as a Relationship/Membership Stream (RMS) — a key constituent component of the IUC that is summarized in" Figure 22 below [1st DOI of Table VIII]:

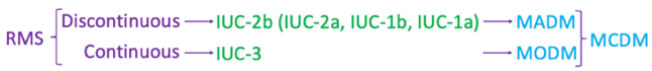


Figure 22. RMS Paradigms for the IUC [1st DOI of Table VIII]

The following Sections IIIC1 through IIIC6 stem from the 4th DOI of Table IV.

1) Nondominated Solution (NS)

Wu had noted the opportunity of transforming FN-related Fuzzy Optimization and Decision Making (FODM) problems to "Scalar Optimization Problem[s]" (SOPs), which can be efficiently resolved to segue to the nondominated solution, wherein "no one objective function can be improved without" a concurrent degradation to "the other objectives," and "the OSNS can then be ascertained." [3rd DOI of Table IV][105].

2) Optimal Corresponding (OC) GLfSFN-based membership function

Interestingly, "regarding 'best-fix approximation[s],' Lakshmana had reported on the efficacy of the "approximations of general non-linear FNs' by way of higher-order linearized Generalized 'f'-gonal FN/SFN forms, 'such as *Tr*iangular, *Tra*pezoidal,' as well as *Pe*ntagonal, *He*xagonal, *Hep*tagonal, *O*ctago*n*al, etc." [3rd DOI of Table IV]; "these can be re-expressed as "G*Tr*FN, G*Tp*FN, G*Pe*FN, G*Hx*FN, G*Hp*FN, G*On*FN, etc., respectively" [3rd DOI of Table IV]. "According to Velu and Ramalingam, 'best-fit approximations' can be improved 'when higher-order piecewise linear' FNs are utilized to approximate 'non-linear information'" [3rd DOI of Table IV][106]. "Along this vein, Augustin asserts that, as one example, G*Hp*FN 'can represent more intricate and nuanced degrees of uncertainty' since 'certain apropos 'f'-gonal FN/SFN forms' are quite good at 'preserving ambiguity'" [3rd DOI of Table IV][107]. "Ban, another advocate of this principle" is satisfied even with "*Tr*iangular, *Tra*pezoidal, and semi-*Tra*pezoidal for the 'preserv[ing]…and weight[ing]' of ambiguity" [3rd DOI of Table IV][108]. "The pathways for deriving the OSNS (a Multi-Objective Decision Making or MODM problem) and the selection of the 'f'-gonal FN/SFN form (a FODM and MCDM problem) are informed by the ICSM2."

3) Preferred OCGLfSFN

"A goal of the involved [Metaheuristic Algorithm] MA (for which the MAM is responsible) is to ascertain the OCGL*f*SFN. Prior to segueing to this OC form, there is a Precursor (P-) non-OC form (i.e., P-GL*f*SFN); for example, Augustin acknowledged…G*Hp*FN for its ability to 'represent more intricate and nuanced degrees of uncertainty' while Ban favored G*Tr*FN, G*Tp*FN, and semi-G*Tp*FN for the preservation of ambiguity and weighted ambiguity" [3rd DOI of Table IV][116][117]. "Whatever the preferred form, the choice of the precursor Generalized 'f'-gonal FN/SFN form (e.g., P-GL*f*SFN) affects the efficacy of the utilized 'defuzzification method' (i.e., 'the transformation of a FN/SFN into a crisp form')" [3rd DOI of Table IV]. "The significance of this center upon the intricacy that 'as the LHM contends with the counterpoising of' A/U, the precursor non-OC form, which best preserves ambiguity, is likely to be optimal for facilitating/deriving the OCGLfSFN."

4) The Ranking of FNs/SFNs

Significantly, "there are numerous 'ranking methods for the discussed pre-cursor [non-OC form] Generalized 'f'-gonal FN/SFN form (e.g., P-GL*f*SFN), and the appropriate selection" is central" [3rd DOI of Table IV]. For example, Velu and Ramalingam noted that 'a ranking method which works very well for' G *He*xagonal FNs/SFNs 'may have some shortfalls when it is extended for' G *O*ctago*n*al FNs/SFNs" [3rd DOI of Table IV][105]. "Similarly, 'a ranking method which works very well for' G *O*ctago*n*al FNs/SFNs might have 'some shortfalls when it is used for' *Tr*iangular or *Tra*pezoidal FNs/SFNs" [3rd DOI of Table IV][109]. "In any case, the ranking mechanism (facilitated by the ACWS) informs the precursor non-OC to final OC form."

5) Similarity Measure (SimM) Challenge for FN/SFN

With regards ranking methods, the underpinning measures typically involve various SimMs. Gogoi & Chutia noted that while there are a myriad of methods (each with advantages/drawbacks), a universally accepted 'silver bullet' SimM for ascertaining the similarity between FNs/SFNs 'does not necessarily exist'" [3rd DOI of Table IV][110]. "They also noted that a 'literature survey reveals that most of the' SimM 'are being developed based upon' the following parameters: "geometric distances, height, area, perimeter, 'Center of Gravity (COG),' 'Radius of Gyration (ROG),' etc." [3rd DOI of Table IV][110]. It was noted in [110] that for various studies, with the exception 'of Hejazi et al. (2011),' certain 'glass ceiling' SimM methods (e.g., "'failing to give reasonable similarity between pairs of FNs when one FN is identical for both the pairs) '*are being carried forward' into contemporary works*" [3rd DOI of Table IV][110]. This is reminiscent of our prior finding that certain bugs/issues in various frameworks/libraries/toolkits, such as made available via assorted developer platforms, were being carried forward into various projects/papers. To aggravate matters, FNs are simply a special case of a FS, and 'beyond FS, there' are other FS variations; these include the IFS, Pythagorean Fuzzy Set (PFS), and Neutrosophic Fuzzy Set (NFS)" [3rd DOI of Table IV][111][112]. "The IFS, which is often leveraged for 'coalition decision-making,' is comprised of constituent elements that 'have both membership function u and non-membership function v, such that $u + v <=1$, and hesitation margin h, such that $u +$

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

$v + h = 1$'" [3rd DOI of Table IV]. "Other situations are better addressed by PFS, 'wherein $u + v >=1$ (or $u + v <=1$) and $u^2 + v^2 + h^2 = 1$'" [3rd DOI of Table IV][111]. "Yet other cases are better handled by NFS, which combines "FS with NS" [3rd DOI of Table IV][112]; "delving into this, Das notes that while FS addresses "uncertainty" by the utilization of 'membership grade,' Smarandache's NS tackles 'uncertainty using truth, indeterminacy, and falsity member grades'" [3rd DOI of Table IV][112]. "Furthermore, Ashraf, Gundogdu & Kahraman, Mahmood, etc. have 'contributed to the [overall] notion of…SFS, which is the generalized structure over' the referenced FS (e.g., IFS, PFS, and NFS)" [3rd DOI of Table IV][113].

6) SimMs and Distance Measures (DMs) for SFS/T-SFS

"Various SimM approaches have been adapted for the SFS ecosystem, as noted by Zhang, and Wei observes, by way of example, that a plethora of 'SimMs for SFS based on the cosine and cotangent function have been' put forth" [3rd DOI of Table IV][114]. "Likewise, certain combinatorial, such as 'Jaccard, Exponential, Square root cosine for SFS,' etc., have been employed as pragmatic implementations of SimMs" [3rd DOI of Table IV][115]. "With regards to DMs, Donyatalab and others have examined 'Minkowski, Minkowski-Hausdorff, Weighted Minkowski and Weighted Minkowski-Hausdorff distances for SFSs'" [3rd DOI of Table IV]. "Overall, there have been numerous SimM and DM advances, and among these, researchers, such as Wu, have 'focused on the T-SFS,' which is a 'specific case of NS' (a.k.a., 'n-hyper SFS')" [3rd DOI of Table IV][114][115]. According to Wu, T-SFS is quite adept in contending with "uncertainty information" and "can handle information that SFSs…cannot process" [3rd DOI of Table IV][115]. Accordingly, the SimMs/DMs of T-SFS show promise for higher efficacy.

IV. CONCLUDING REMARKS

For a CAA to successfully operationalize a high efficacy NLG and engage in unscripted dialogues, it needs to have RDM; for a CAA to learn/tune from each and every engagement, it needs to not only have an efficient ML mechanism, but it also needs to have a high performance ML2 mechanism. With these two facets in hand, the CAA is more likely to be able to engage in real-time conversation without any undesired delays in performance. The benchmark as to whether the aforementioned is successfully actualized is predicated upon whether the CAA responses are "consistent, coherent, and valid." As noted in Section I, this is non-trivial to achieve. Delving deeper (from the high-level references to NLG and RDM), it was necessary, for the purposes of this paper, to first address SDM (specifically, SDM-A/U for RWS).

While the longer-term impetus of this paper is to address SDM-A/U for AIS/AICDS/AS (e.g., AUVs, USVs, UAVs, etc.) as well as other applications of SAE 5, the proxy application of CAIR was selected to better understand the multi-stage challenges of SDM-A/U. For SDM-A/U, at each DM point, a choice needs to be made as to whether a "full resolution," "partial resolution," or "no resolution" action will be taken. In contrast to a single-stage DM, a key challenge of SDM is to optimize the overall outcome over the full range of multi-stage DM points.

One approach, among others, for tackling the SDM-A/U challenge resides in the leveraging of an LHM, which carefully considers – concurrently – the notions of A/U. The LHM is powered by, among other key modules, an ICSM2 and a MAM; for the purposes of the CAA use case and CAIR application discussed in this paper, a CAIR-A was added as well. The ICSM2 starts with an initial set of considerations and ascertainments (UnS, EqS, EquivS, OrS, POSET, UnS-Iso, POSET-Iso, IV, etc.) prior to proceeding to an IsoP, whose computational complexities can vary greatly. Of course, the pre-IsoP approach is quickly handled while the IsoP determination and approach is also contingent upon how much time is available. For example, while a full IsoP approach might be selected under UDC, a relaxed IsoP (r-IsoP) might be opted for under CDC or in-between CDC and UDC. The choice of approach is heavily contingent upon the CW mechanism of the ACWS, which remains a central element of not only the entire described apparatus, but also both ICSM2 and MAM operations in particular.

There exists a mutual reinforcement interplay between ACWS and MMSO, and the overarching MCDM rubric (which tends to have a myriad of conflicting objectives or criteria, particularly for RWS) can be contextualized by a non-dominated solution (which illuminates a number of best-scenario trade-offs; however, it does not indicate a single optimal solution). It had previously been determined that OSNS (particularly the extended OSONS version) can be useful in this regard, and it has been found to be of value-added proposition for AIS not only in the area of better understanding trade-offs but has also been acknowledged for its value in the areas of I&E (under the STEA rubric) and feature importance analysis. For this extrapolated journal paper, it should be remembered from [1] varied SHAP approaches differ in their local and global efficacies. It should also be remembered that CW/ACWS ranks the criteria (i.e., CW/ACWS prioritizes what is most important before the alternatives are even reviewed), wherein criteria are the factors utilized to assess, compare, and contrast the set of alternatives. Each of the alternatives are characterized by features. Interestingly, IbSOA can leverage the clustering of features into coalitions to facilitate computational efficiency. Section IB2 reviews the notion of "updating a heuristic" and notes that, contrary to popular opinion, heuristics may segue to sub-optimal paradigms more frequently than anticipated; in essence, the notion of simply using a heuristic as an accelerant is far too simplistic and may be disappointing. Of significance, the "updating of a heuristic" may necessitate sufficient STEA (and IbSOA may be useful in this regard) so as to facilitate ML2; overall, heuristics for RWS should be able to: (1)self-recognize that the heuristic is not qualified to handle the incoming information and needs to refer the matter back to the higher-order hyper-heuristic, (2) mitigate against the ongoing

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

prospective brittleness of the heuristic itself, and (3) have the capability to operationalize the involved [meta]heuristic-algorithm amalgam in an advantageous fashion. Section IB2 culminates with the overarching equation of this paper: ASL-IMC⇢RMMI, wherein ASL equates to AIS⇢SDM⇢LHM, IMC equates to ICSM2+MAM+CAIR-A, and RMMI equates to RDM ⇢MR/NMR⇢MNTZ⇢IbSOA. Of course, IbSOA is captured within the first part of the title of this paper: "Interstitial b-SHAP-Owen Amalgam."

IbSOA is put forth as a central item for the second part of the title of this paper: "Enhancement of Artificial Intelligence System-Centric Sequential Decision-Making," and Section II had opened with the notion that "the AIS approach towards DMP, such as SDM, may involve a variety of" RMs. Each of the varied RMs has certain monotonic and/or non-monotonic tendencies, and depending upon a paradigm of UDC or CDC, the MAM may opt for certain RM pathways and proceed in stages depending upon the time remaining and/or any time adjustments that may occur along the SDM pathway. The MAM supports the ICSM2, particularly when it is determined that IsoP needs to occur. Along this vein, the ICSM2 operationalizes the IsoP comparison via the use of GFN/GFS-IUC and Paradoxes-A/U. By working in conjunction, the ICSM2 and MAM combinatorial form a powerful engine for LHM. The addition of the CAIR-A module enhances the combinatorial by serving as a precursor function; the CAIR-A endeavors to ascertain what primary, secondary, etc. RMs might be involved. The involved CAIR-related CW/ACWS-MMSO pre-sort tends to be of value-added proposition for RWS SDM paradigms. The "A" facet of the CAIR-A Module is an orchestration mechanism that leverages the MAM (which supports the LHM) in conjunction with a Birnbaum Importance assignment; it should be remembered that the Birnbaum importance measure is often leveraged as a Lower-Level Heuristic (LLH) for use in resolving Component Assignment Problem (CAP) (which centers upon optimally placing components at various positions within an AIS to maximize reliability).

To clarify this point, if N+V in it of itself does not suffice (and IsoP is required), then there is an accompanying increased MR/NMR skewing risk that should be mitigated. In particular, as pertains to the MNTZ, it has been noted consistently throughout this paper and the specified apriori examinations that there exists a tendency for spawning to the NP-Hard side (e.g., non-monotonic, non-polynomial, and even non-continuous functions). Roughly speaking, the problem with this is that LLMs struggle with emulating human behavior on the non-monotonic side, and oftentimes, the AIS (in this case, the CAA) will take the opposite stance as what the human would choose. Cognizant of the range of paradoxes (e.g., Ellsberg, Machina, Allais, etc.) that influence this paradigm, due consideration should be paid regarding the associated preference ranking, and the related issue of RR has been previously scrutinized. While it seems that focusing on certain RMs (e.g., deductive) might seem computationally more tractable, it would preclude more RWS RM approaches (e.g., probabilistic, temporal, and graph-based). Yet these parenthetical approaches, among others, can segue to non-monotonic unexpectedly. Moreover, the desired related CBR and inductive approaches are at much higher risk for segueing to non-monotonic. Finally, some of the other desired more sophisticated approaches (e.g., analogical, abductive) squarely reside in the non-monotonic. Depending upon the RM pathway opted for, the related DM will likely impact the downstream behavior within the MNTZ in terms of a skewing towards monotonic or non-monotonic behavior. In essence, while the gamut of isomorphic cases can be construed to be analogies, not all analogies are construed to be isomorphic. This particular determination is central because if IsoP can be avoided, then the expenditure of an unknown amount of computation resources can by bypassed.

The significance of the IbSOA is that, fundamentally, the referenced Owen value nicely captures the intricate interactions between/among the constituent members of the involved feature set and, as noted in [1], "extends the Shapley value (which well captures the individual feature contributions) in a consistent fashion." These pertinent insights into CAA behavior at the MNTZ, as noted by [2], "can potentially be quite meaningful for elevating CAIR-related coherence and consistency (with the concomitant validity)." After all, as [2] further articulates, "maintaining coherence and monotonicity is non-trivial, as the involved AIS might discern connections (particularly those that are non-monotonic) within the evolving dataset. In the context of CAIR, non-monotonic aspects can arise as incoming information can re-contextualize and/or contradict matters."

As noted in [2], the effectiveness of particular RWS implementations, such as that of CAA, "is often predicated upon consistency and reliability" (along with the associated notions of coherency and validity), and this aspect is the essence of CAIR. As noted by [2], the CAA should, ideally, provide 'human-like conversations' by comprehending user intent, maintaining context, and putting forth pertinent responses' consistent with 'the principle of CAIR;' in essence, a key tenet of CAIR is that responses provided by the CAA remain steady in their validity not only during the course of the involved interaction, but also over the course of the multi-turn conversation. In summary, this paper posits that the CAIR-A notion isof VAP to the ongoing challenge of conversational coherence and RDM. Future work will involve more quantitative experimentation, and as discussed in Section IIB, the issues of provenance/pedigree, as pertains to RM selection, regarding the "who" and the "where" (as contrasted to simply the "what" and "when" the information is made available) will be further explored.

REFERENCES

[1] S. Chan, "Interstitial b-SHAP-centric Amalgam for the Enhancement of an AI-centric Construct Validity Approach," *The Seventeenth International Conference on Future Computational Technologies and Applications (Future Computing 2025)*, Apr. 2025, pp. 19-26.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[2] S. Chan, "A Prospective Monotonic/Non-Monotonic Transition Zone Impediment for Concept Model-Centric Artificial Intelligence Systems," *The Second International Conference on AI-based Systems and Services (AISyS 2025)*, in press.

[3] C. Wanke and D. Greenbaum, "Sequential Congestion Management with Weather Forecast Uncertainty," *American Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation and Control Conference and Exhibit*, Aug. 2008, pp. 1-19.

[4] E. Keogh and A. Mueen, "Curse of Dimensionality," in *Encyclopedia of Machine Learning*, Boston, MA. Springer, 2011, pp. 257-258.

[5] S. Chan, "Resilient Decision Systems and Methods," U.S. 11,862,977, Jan. 2, 2024.

[6] S. Chan, "Resilient Decision Systems and Methods," U.S. 12,362,565, Jul. 15, 2025.

[7] N. Lemons, B. Hu, W. Hlavacek, "Hierarchical graphs for rule-based modeling of biochemical systems," *BMC Bioinformatics*, vol. 12, pp. 1-13, Feb. 2011.

[8] E. Firt, "Analogical reasoning as a core AGI capability," in *AI and Ethics*, vol. 5, pp. 1-15, Jul. 2025.

[9] D. Gentner and L. Smith, "Analogical Reasoning," in *Encyclopedia of Human Behavior*, Oxford, UK. Elsevier, 2012, pp. 130-136.

[10] J. Thibaut, Y. Glady, and R. French, "Understanding the What and When of Analogical Reasoning Across Analogy Formats: An Eye-Tracking and Machine Learning Approach," *Cogn. Sci*., vol. 46, pp. 1-41, Nov. 2022.

[11] G. Nandi, "Probabilistic Reasoning," in *Principles of Soft Computing Using Python Programming: Learn How to Deploy Soft Computing Models in Real World Applications*, IEEE, 2024, pp.159-196.

[12] A. Nafar, K. Venable, and P. Kordjamshidi, "Reasoning over Uncertain Text by Generative Large Language Models," *Thirty-Ninth Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*, Jan. 2025. pp. 1-21.

[13] Leeuwenberg and M. Moens, "A Survey on Temporal Reasoning for Temporal Information Extraction from Text," *J. of Artifical Intelligence Research*, vol. 66, pp. 341-380, Sep. 2019.

[14] S. Xiong, A. Payani, R. Kompella, and F. Fekri, "Large Language Models Can Learn Temporal Reasoning," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 10452-10470, Aug. 2024.

[15] S. Grote-Garcia, "Deductive Reasoning," in *Encyclopedia of Child Behavior and Development*. Boston, MA. Springer, 2011, pp. 477-478.

[16] "Deductive Reasoning," Taylor & Francis. [Online]. Accessed: Mar. 1, 2025. Available: https://taylorandfrancis.com/knowledge/Engineering_and_technology/Engineering_support_and_special_topics/Deductive_reasoning/

[17] "Inductive Reasoning," University of Illinois Sprinfield, ION Professional eLearning Programs, [Online]. Accessed: Mar. 1, 2025. Available: https://www.uis.edu/ion/resources/oiai/inductive-reasoning

[18] A. Sandoval-Hernandez and D. Rutkowski, "Embracing complexity: abductive reasoning as a versatile tool for analyzing international large-scale assessments," in *Educ. Assessment Eval. And Accountability*, vol. 37, pp. 255-271, Dec. 2024.

[19] P. Thagard and C. Shelley, "Abductive Reasoning: Logic, Visual Thinking, and Coherence," in *Logic and Scientific Methods*, vol. 259, pp. 413-427, 1997.

[20] A. Belzen, P. Engelschalt, and D. Kruger, "Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence," *Educ. Sci.,* vol. 11, pp. 1-11, Sep. 2021.

[21] A. Yan and Z. Cheng, "A Review of the Development and Future Challenges of Case-Based Reasoning," *Appl. Sci.*, vol. 14, pp. 1-22, Aug. 2024.

[22] "Case-based reasoning," *Taylor & Francis*, [Online]. Accessed: Mar. 1, 2025. Available: https://taylorandfrancis.com/knowledge/Engineering_and_technology/Artificial_intelligence/Case-based_reasoning/

[23] S. Chan, "Enhanced Robust Convex Relaxation Framework for Optimal Controllability of Certain Large Complex Networked Systems," *The 2022 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications (IARIA Congress 2022)*, Jul. 2022, pp. 87-96.

[24] L. Cao, "GraphReason: Enhancing Reasoning Capabilities of Large Language Models through A Graph-Based Verification Approach," *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations,* Aug. 2024, pp. 1-12.

[25] "Athina-ai/athina-evals," *Github*, [Online]. Accessed: Mar. 1, 2025. Available: https://github.com/athina-ai/athina-evals/blob/main/examples/conversation_coherence.ipynb

[26] G. Harman, "The Inference to the Best Explanation" *Philosophical Review*, vol. 74, pp. 323-327, Jan. 1965.

[27] K. Ferguson, "Monotonicity in Practical Reasoning," *Argumentation*, vol. 27, pp. 335-346, Sep. 2003.

[28] A. Bundy and L. Wallen, "Non-Monotonic Reasoning," in *Catalogue of Artificial Intelligence Tools*, Berlin, Germany. Springer Nature, 1984, pp. 83.

[29] A. Fuhrmann, "Non-Monotonic Logic," in *Routledge Encyclopedia of Philosophy*, Routledge, 1998.

[30] P. Hentenryck, "Abduction and Abductive Logic Programming," in *Logic Programming: The 11th International Conference*, MIT Press, 1994, pp.18-19.

[31] G. Paul, "Approaches to abductive reasoning: an overview," *Artificial Intelligence Review*, vol. 7, pp. 109-152, Apr. 1993.

[32] V. Lagerkvist, M. Maizia, and J. Schmidt, "A Fine-Grained Complexity View on Propositional Abduction – Algorithms and Lower Bounds," Jun. 2025. [Online]. Accessed: Mar. 1, 2025. Available: https://www.arxiv.org/abs/2505.10201

[33] A. Leidinger, R. Rooij, and E. Shutova, "Are LLMs classical or nonmonotonic reasoners? Lessons from generics," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 558-573, Aug. 2024.

[34] S. Han, K. Ransom, A. Perfors, and C. Kemp, "Inductive reasoning in humans and large language models" in *Cognitive Systems Research*, vol. 83, pp. 1-28, Jan. 2024.

[35] Mehran Kazemi, et al., "BoardgameQA: A Dataset for Natural Language Reasoning with Contradictory Information," *Proceedings of the 37th Conference on Neural Information Processing Systems (NIPS '23)*, Dec 23, pp. 39052-39074.

[36] K. Jantke, "Monotonic and non-monotonic inductive inference," *New Generating Computing*, vol. 8, pp. 349-3601, Feb. 1991.

[37] K. Jantke, "Monotonic and non-monotonic inductive inference of functions and patterns," *Lecture Notes in Computer Science*, vol. 543, pp. 161-177, Jan. 2005

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[38] H. Mirtagioglu and M. Mendes, "On Monotonic Relationships," *Biostatistics and Biometrics*, vol. 10, pp. 1-11, May 2022.

[39] M. Kerber and E. Melis, "Two kinds of non-monotonic analogical inference," *Lecture Notes in Computer Science*, vol 1085, pp. 361-374, Jan. 2005.

[40] X. Liu, "Probabilistic approaches to non-monotonic reasoning," *Proceedings of the Nineteenth International Symposium on Multiple-Valued Logic*, May 1989, pp. 378-382.

[41] B. Grosof, "Non-Monotonicity in Probabilistic Reasoning," Machine Intelligence and Pattern Recognition," vol. 5, pp. 237-249, 1988.

[42] C. Baral and J. Zhao, "Non-monotonic Temporal Logics for Goal Specification," *Proceedings of the 20th International Joint Conference on Artifical intelligence (IJCAI '07)*, Jan. 2007, pp. 236-242.

[43] A. Hunter, "Non-monotonic Reasoning in Deductive Argumentation," pp. 1-24, Sep. 2018, [Online]. Accessed: Mar. 1, 2025. Available: https://arxiv.org/abs/1809.00858.

[44] L. Pereira and A. Nerode, "An assumption-based framework for non-monotonic reasoning," in *Logic Programming and Non-Monotonic Reasoning: Proceedings of the Second International Workshop*, MIT Press, 1993, pp.171-189.

[45] G. Paulino-Passos and F. Toni, "Monotonicity and Noise-Tolerance in Case-Based Reasoning with Abstract Argumentation," *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning*, Nov. 2021, pp. 508-518.

[46] A. Bochman, "The Many Valued and Nonmonotonic Turn in Logic," in *Handbook of the History of Logic*, vol. 8, pp. 13-689, 2007.

[47] C. Wearing, "Allegory, Metaphor, and Analogy," *Journal of Pragmatics*, vol. 202, pp. 66-79, Dec. 2022.

[48] R. Holme, "Allegory and Analogy: Teaching with Extended Metaphors," in: *Mind, Metaphor and Language Teaching*, pp. 98-119, 2004.

[49] N. Burton, "The Psychology of Allegory and Metaphor," *Psychology Today*, Apr 2021. [Online]. Accessed: Mar. 1, 2025. Available: https://www.psychologytoday.com/us/blog/hide-and-seek/202104/the-psychology-allegory-and-metaphor.

[50] D. Hofstadter, "Epilogue: Analogy as the Core of Cognition," in *The Analogical Mind*, 2001.

[51] K. Holyoak, D. Gentner, and B. Kokinov, "Introduction: The Place of Analogy in Cognition," in *The Analogical Mind*, 2001.

[52] D. Hofstadter and E. Sander, "Surfaces and Essences: Analogy as the fuel and fire of thinking," *APA PsycInfo*, 2013.

[53] T. Wijesiriwardene, A. Sheth, V. Shalin, and A. Das, "Why Do We Need Neurosymbolic AI to Model Pragmatic Analogies?" *IEEE Intelligent Systems*, vol. 38, pp. 12-16, Sep. 2023.

[54] Chakraborty, Mondal, Alam, Ahmadian, Senu, De, and Salahshour, "The Pentagonal Fuzzy Number: Its Different Representations, Properties, Ranking, Defuzzification and Application in Game problems," *Symmetry*, vol. 11, pp. 1-29, Feb. 2019.

[55] A. Mert, "Defuzzification of Non-Linear Pentagonal Intuitionistic Fuzzy Numbers and Application in the Minimum Spanning Tree Problem," *Symmetry*, vol. 15, pp. 1-29, Oct. 2023.

[56] I. Couso and H. Bustince, "From fuzzy sets to interval-valued and Atanassov intuitionistic fuzzy sets: a unified view of different axiomatic measures," *IEEE Transactions on Fuzzy Systems*, vol. 27, pp. 362-371, Jul. 2018.

[57] M. Luo, W. Li, and H. Shi, "The Relationship between Fuzzy Reasoning Methods Based on Intuitionistic Fuzzy Sets and Interval-Valued Fuzzy Sets," *Axioms*, pp. 1-13, Aug. 2022.

[58] R. Verma and S. Chandra, "Interval-Valued Intuitionistic Fuzzy-Analytic Hierarchy Process for evaluating the impact of security attributes in Fog based Internet of Things paradigm," *Computer Communications*, vol. 175, pp. 35-46, Jul. 2021.

[59] A. Lu and W. Ng, "Vague Sets or Intuitionistic Fuzzy Sets for Handling Vague Data: Which One is Better?" *Lecture Notes in Computer Science*, vol. 3716, pp. 401-416, Oct. 2005.

[60] S. Alkhazaleh, "Neutrosophic Vague Set Theory," *Critical Review*, volume 10, pp. 29-39, Jan. 2015.

[61] A. Khuman, "The similarities and divergences between grey and fuzzy theory," *Expert Systems with Applications*, vol. 186, Dec. 2021.

[62] Sifeng, L., Forrest, J., and Yingjie, Y. (2011). "A brief introduction to grey systems theory," in *Grey Systems and Intelligent Services (GSIS)*, pp. 1–9, 2011.

[63] M. Shaker and M. Moore-Clingenpeel, "The known knowns, known unknowns, and unknown unknowns of surveys and sleep," *Ann. Allergy Asthma Immunol*., vol 129, pp. 669-670, Dec. 2022.

[64] S. McGregor, "Learning with Donald Rumsfeld – flexible learning: the relevance and resonance of multiprofessional learning in primary care," *Br. J. Gen Pract*., vol. 54, pp. 722-723, Sep. 2004.

[65] M. Lang, "First-Order and Second-Order Ambiguity Aversion," *Management Science*, vol. 63, Jun. 2016.

[66] M. Machina and M. Siniscalchi, "Chapter 13 – Ambiguity and Ambiguity Aversion," in *Handbook of the Economics of Risk and Uncertainty*," vol. 1, pp. 729-807, 2014.

[67] R. Jia, E. Furlong, S. Gao, L. Santos, I. Levy, "Learning about the Ellsberg Paradox reduces, but does not abolish, ambiguity aversion," *PLOS One*, pp. 1-24, Mar. 2020

[68] T. Coleman, "Probability, Expected Utility, and the Ellsberg Paradox," Feb 2011. [Online]. Accessed: Mar. 1, 2025. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1770629

[69] D. Ellsberg, "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics*, vol. 75, pp. 643-669, Nov. 1961.

[70] D. Chen, "Is ambiguity aversion a preference? Ambiguity aversion without symmetric information," *Journal of Behavioral and Experimental Economics*, vol. 111, Aug. 2024.

[71] B. Weber and W. Tan, "Ambiguity aversion in a delay analogue of the Ellsberg Paradox," *Judgment Decision Making*, vol. 7, pp. 383-389, July 2012.

[72] K. Binmore, L. Stewart, A. Voorhoeve, "How Much Ambiguity Aversion? Finding Indifferences between Ellsberg's Risky and Ambiguous Bets," *Journal of Risk and Uncertainty*, vol. 45, pp. 215-238, Nov. 2012.

[73] M. Leopold, "Smarter in the Long-Term: Diminishing Ambiguity Aversion in a Repeated Ellsberg Urn Task," [Online]. Accessed: Mar. 1, 2025. Available: https://cogsci.yale.edu/sites/default/files/files/Thesis2016Leopold.pdf

[74] D. Chen, "Is ambiguity aversion a preference? Ambiguity aversion without asymmetric information," *Journal of Behavioral and Experimental Economics*, vol. 111, Aug. pp. 1-47, 2024.

[75] Y. Halevy and V. Feltkamp, "A Bayesian Approach to Uncertainty Aversion," *Review of Economic Studies*, vol. 72, pp. 449-466, Apr. 2005.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[76] U. Segal, "The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach," *International Economic Review*, vol. 28, pp. 175-202, Feb. 1987.

[77] J. Moreno-Jimenez and L. Vargas, "A Model to Deal with Uncertainty in Ellsberg's Paradox: The Analytic Hierarchy Process with Feedback," *The International Symposium on the Analytic Hierarcy Process*," Aug. 1996, pp. 1-12.

[78] B. Jabarian and. S. Lazrus, "A Two-Ball Ellsberg Paradox," CESifo Working Paper No. 10745, pp. 1-54, Nov. 2023. [Online]. Accessed: Mar. 1, 2025. Available:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4636003

[79] D. Aerts, S. Sozzo, and J. Tapia, "A Quantum Model for the Ellsberg and Machina Paradoxes," [Online]. Accessed: Mar. 1, 2025. Available:https://arxiv.org/abs/1208.2354.

[80] J. Quiggin, "Subjective utility, anticipated utility, and the Allais paradox," *Organizational Behavior and Human Decision Processes*, vol. 35, pp. 94-101, Feb. 1985.

[81] S. Ferrari-Toniolo, L. Seak, and W. Schultz, "Risky choice: Probability weighting explains independence axiom violations in monkeys," *J. Risk Uncertain.*, vol. 65, pp. 319-351, Jul. 2022.

[82] P. Blavatskyy, "Probabilistic independence axiom," *The Geneva Risk and Insurance Review*, vol. 46, pp. 21-34, Jan 2020.

[83] J. Jim, "AI preference prediction and policy making," *AI & Soc*., pp. 1-15, Jul. 2025.

[84] A. Aldoseri, "Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges," *Appl. Sci*., vol. 13, pp. 1-33, Jun. 2023.

[85] A. Chen et al., "Preference Learning Algorithms Do Not Learn Preference Rankings," *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Jun. 2025, pp. 101928-101968.

[86] L. Qin, Y. Zhu, S. Liu, X. Zhang, and Y. Zhao, "The Shapley Value in Data Science: Advances in Computation, Extensions, and Applications," *Mathematics*, vol. 13, pp. 1-21, May 2025.

[87] B. Rozemberczki, "The Shapley Value in Machine Learning," *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, Jul. 2022, pp. 5572-5579.

[88] A. Casjus, "The shapley value, the owen value, and the veil of ignorance," *Int. Game Theor. Rev*., vol. 11, pp. 453-457, Dec. 2009.

[89] S. Lopez, "On the relationship between Shapley and Owen values," *Central European J. of Oper. Res*., vol. 17, pp. 415-423, Dec. 2009.

[90] S. Beal, M. Diss, and R. Takeng, "New axiomatisations of the Diversity Owen and Shapley values," *CRESE Working Paper 2024-09*, Feb. 2024. [Online]. Accessed: Mar. 1, 2025. Available: https://hal.science/hal-04502031v1/document.

[91] Z. Wu, "Evaluation of Provincial Economic Resilience in China Based on the TOPSIS-XGBoost-SHAP Model," *J. of Math.*, pp. 1-12, Oct. 2023.

[92] X. Wang and Z. Piao, "An interpretable dynamic risk assessment approach in insurance: integrating TOPSIS, GM (1,1) and SHAP," *Proc. of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area Int. Conf. on Digit. Econ. and Artif. Intell. (DEAI)*, pp. 697-701, Jul. 2024.

[93] S. Chan, "Mitigation Factors for Multi-domain Resilient Networked Distributed Tessellation Communications," *Fifth Int. Conf. on Cyber-Technol. and Cyber-Syst. (CYBER 2020)*, Aug. 2020, p. 66-73.

[94] E. Borgonovo, E. Plischke, and G. Rabitti, "The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective," *European J. of Oper. Res.*, vol. 318, pp. 911-926, Nov. 2024.

[95] M. Brans and B. Mareschal, "Promethee Methods," *Multiple Criteria Decision Analysis: State of the Art Surveys*, pp. 163-186, Jan. 2005.

[96] J. Brans and P. Vincke, "A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making), *Manage. Sci*., vol. 31, pp. 647-656, Jun. 1985.

[97] Y. Lin and T. Chen, "Type-II fuzzy approach with explainable artificial intelligence for nature-based leisure travel destination selection amid the COVID-19 pandemic," *Digit. Health,* vol. 8., pp. 1-15, Jun. 2022.

[98] M. Aazadfallah, "A New Feature of Rank Reversal in Some of MADM Models," *J. of Appl. Inf. Sci*., vol. 4, pp. 1-11, 2015.

[99] M. Mase, A. Owen, and B. Seiler, "Explaining black box decision by Shapley cohort refinement," Oct. 2020. [Online]. Accessed: Mar. 1, 2025. Available: https://arxiv.org/pdf/1911.00467.

[100] L. Mancinska, D. Roberson, R. Samal, S. Severini, and A. Varvitsiotis, "Relaxations of Graph Isomorphism," *Cambridge Quantum*, pp. 1-15, Jul. 2017.

[101] A. Atserias, "On Continuous and Combinatorial Relaxations of Graph Isomorphism," [Online]. Accessed: Mar. 1, 2025. Available: https://www.cs.upc.edu/~atserias/talks/graphisotalk/graphisotalk.pdf

[102] Y. Aflalo, A. Bronstein, and R. Kimmel, "On convex relaxation of graph isomorphism," *Proc. Natl. Acad Sci. (PNAS),* vol. 112, pp. 1-6, Mar. 2015, pp. 2942-2947.

[103] A. S. Khuman, Y. Yang, R. John, and S. Liu, "R-fuzzy sets and grey system theory," in *Proc. 2016 IEEE Int. Conf. Syst. Man Cybernetics*, 2016, pp. 004555–004560.

[104] A. S. Khuman, "The similarities and divergences between grey and fuzzy theory," *Expert Syst. Appl.*, vol. 186, pp. 1–11, Dec. 2021.

[105] S. Nayak, "Multiobjective Optimization," Fundamentals of Optimizaiton Techniques with Algorithms, pp. 253-270, 2020.

[106] L. Velu and B. Ramalingam, "Total Ordering on Generalized 'n' Gonal Linear Fuzzy Numbers," *Int. J. of Comput. Intell. Syst*., vol. 16, pp. 1-19, Feb 2023.

[107] E. Natarajan, F. Augustin, M. Kaabar, C. Kenneth, and K. Yenoke, "Various defuzzification and ranking techniques for the heptagonal fuzzy number to prioritize the vulnerable countries of stroke disease," *Results in Control and Optim*., vol. 12, pp. 1-30, Sep 2023.

[108] V. Nayagam and J. Murugan, "Hexagonal fuzzy approximation of fuzzy numbers and its applications in MCDM," *Complex & Intell. Syst.*, vol. 7, pp. 1459-1487, Feb 2021.

[109] L. Velu and B. Ramalingam, "Total Ordering on Generalized 'n' Gonal Linear Fuzzy Numbers," *Int. J. of Comput. Intell. Syst*., vol. 16, pp. 1-19, Feb 2023.

[110] M. Gogoi and R. Chutia, "Fuzzy risk analysis based on a similarity measure of fuzzy numbers and its application in crop selection," in *Eng. Appl. of Artif. Intell*., vol. 107, pp. xxx, Jan 2022.

[111] P. Ejegwa, "Pythagorean fuzzy set and its application in career placements based on academic performance using max-min-max composition," *Complex & Intell. Syst*., vol. 5., pp. 165-175, Feb 2019.

[112] S. Das, B. Roy, M. Kar, S. Kar, and D. Pamucar, "Neutrosophic fuzzy set and its application in decision

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

making," *J. of Ambient Intell. and Humanized Comput*., vol. 11, pp. 5017-5029, Mar. 2020.

[113] E. Ozceylan, B. Ozkan, M. Kabak, and M. Dagdeviren, "A state-of-the-art survey on spherical fuzzy sets," *J. of Intell. & Fuzzy Syst.: Appl. in Eng. and Technol*., vol. 42, pp. 195-212, Dec 2021.

[114] Y. Donyatalab, F. Gundogdu, F. Farid, S. Seyfi-Shishavan, E. Farrokhizadeh, and C. Kahraman, "Novel spherical fuzzy distance and similarity measures and their applications to medical diagnosis," *Expert Syst. with Appl*., vol. 191, pp. 1-15, Apr 2022.

[115] M. Wu, T. Chen, and J. Fan, "Similarity Measures of T-Spherical Fuzzy Sets Based on the Cosine Function and Their Applications in Pattern Recognition," *IEEE Access*, pp. 98181-98192, May 2020.

[116] E. Natarajan, F. Augustin, M. Kaabar, C. Kenneth, and K. Yenoke, "Various defuzzification and ranking techniques for the heptagonal fuzzy number to prioritize the vulnerable countries of stroke disease," *Results in Control and Optim*., vol. 12, , pp. 1-31, Sep 2023.

[117] V. Nayagam and J. Murugan, "Hexagonal fuzzy approximation of fuzzy numbers and its applications in MCDM," *Complex & Intell. Syst.*, vol. 7, pp. 1459-1487, Feb 2021.

[118] H. Bustince, "Interval-valued Fuzzy Sets in Soft Computing, International Journal of Computational Intelligence Systems, vol. 3, pp. 215-222, June 2010.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Sound and Safety in the Age of Electric Vehicles

Pedestrian Insights from the EVA Project

Flaithrí Neff
The IDEAM Research Institute
Technological University of the Shannon
Limerick, Ireland
e-mail: Flaithri.Neff@tus.ie

Linda Butler
School of Education, Hibernia College
Dublin, Ireland
lbutler@hiberniacollege.net

Dónal Fitzpatrick, Gerald Craddock, Carly Cheevers, Aislinn Conway
The Centre for Excellence in Universal Design
The National Disability Authority
Dublin, Ireland
donal.fitzpatrick@nda.ie
gerald.craddock@nda.ie
carly.cheevers@nda.ie
aislinn.conway@nda.ie

***Abstract* — Electric vehicles (EVs) pose challenges for pedestrian safety due to reduced sound levels at low speeds, prompting regulations mandating Acoustic Vehicle Alerting Systems (AVAS). This article reports findings from a survey of pedestrians with and without vision-impairment. Participants with vision-impairment reported significantly lower perceptions of safety, detectability, and interpretability, while both groups frequently described surprise encounters with EVs. These findings expose three persistent challenges in AVAS regulation: the absence of distance-based design guidance beyond basic sound pressure thresholds; a tendency toward branding-focused rather than functionally informative sound-design; and the uncertain effects of overlapping AVAS signals in future urban soundscapes. More broadly, these issues reflect issues around limited academic engagement in standards and regulatory development, where commercial feasibility has often outweighed perception-research evidence. We argue that future progress depends on embedding pedestrian-centered metrics into standards, allowing psychometric research to inform compliance testing in ways that more accurately reflect real-world pedestrian experience.**

Keywords- Electric Vehicles; Acoustic Vehicle Alerting Systems; Pedestrian Safety; Sound Perception.

I. INTRODUCTION

An electric vehicle (EV) is a mode of transport that is powered by electricity rather than conventional internal combustion fueled by petrol or diesel. It employs an electric motor driven by a rechargeable battery, which can be replenished using dedicated charging stations or an installed home/business electrical outlet. EVs are typically categorized into three main types: 1) Battery Electric Vehicles (BEVs), which are fully electric and contain no internal combustion engine (ICE) components; 2) Plug-in Hybrid Electric Vehicles (PHEVs), which combine an electric motor with a petrol engine, allowing limited electric-only operation before switching to fossil fuel; and 3) Hybrid Electric Vehicles (HEVs), which also include both an electric motor and an ICE, but differ from PHEVs in that the battery is not recharged via an external source but rather through regenerative braking and engine operation.

The primary advantages of EVs include significantly lower tailpipe emissions (in some cases zero), reduced operating costs, lower maintenance requirements due to fewer moving mechanical components, and very quiet operation. It is this significantly-reduced noise output, particularly at low speeds, that has introduced new safety concerns [1][2][3][4] - especially for vision-impaired pedestrians who rely on auditory cues to detect oncoming traffic due to having reduced vision or complete sight-loss. In response, electric and hybrid vehicles are now equipped with an Acoustic Vehicle Alerting System (AVAS), which emits artificial sound to announce the vehicle's presence to nearby pedestrians and cyclists when operating at low speeds (see Section II for information on why AVAS initiates only at low speeds).

Nonetheless, empirical studies consistently show that pedestrians exhibit delayed reaction times and inaccurate distance judgments when engaging with EVs compared to ICE vehicles [5][6][7][8][9][10][11][12]. While AVAS-generated sounds have been shown to improve pedestrian detection rates compared to earlier near-silent EV models, evidence also suggests that current AVAS signatures lack certain acoustic cue-information that is naturally present in conventional ICE vehicle sounds [13]. Therefore, as EV adoption continues to accelerate [14][15], there is a growing imperative for further research into the acoustic characteristics of both ICE acoustics and AVAS, including how they influence pedestrian perception, interaction, and overall safety-outcomes in diverse real-world contexts.

While regulatory frameworks define broad acoustic characteristics and compliance thresholds for AVAS-equipped vehicles, there remains a significant knowledge gap regarding how effective AVAS performs at supporting real-world vehicle localization. Indeed, AVAS standards treat distance as a geometric parameter, but they do not determine whether such distances correspond to pedestrians' perception of a safe margin. In making a small contribution to closing such gaps, we present the EVA (Electric Vehicle Acoustics)

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

study where we set out to gather experiential data from two key pedestrian cohorts: individuals without vision-impairment and those with vision-impairment. Using this comparative approach, our goal is to show how AVAS signals are perceived in practice by pedestrians who can rely on visual cues to compensate for potential AVAS limitations versus those who cannot. This information, therefore, helps evaluate whether current standards and regulations sufficiently account for the needs and experiences of all pedestrians - particularly those with vision-impairments.

In presenting these findings and their implications for standards and policy, the article is organized as follows. Section II introduces the regulatory frameworks that underpin AVAS and Section III explains the methodology adopted for implementing the EVA survey and analyzing its data. Sections IV to VI present the quantitative findings step by step: first a descriptive analysis of the Likert items, then their correlations, and finally the results of significance testing. Section VII turns to the qualitative material, analyzing participants' open-ended responses. Section VIII brings these strands together in discussion, connecting the EVA findings to wider regulatory debates and highlighting the persistent perceptual gaps in AVAS. Section IX sets out the main conclusions, and Section X looks ahead to future work, particularly the development of perception-based metrics, with a final statement on how future AVAS may progress as enabling-technologies come on stream – producing for example context-aware, AI-enhanced AVAS systems.

II. AVAS Regulatory Frameworks

In the European Union, the inclusion of AVAS became a legal requirement for all newly approved electric and hybrid vehicles as of 1 July 2019, under Regulation (EU) 2019/2144 [16]. This regulation amends the earlier Regulation (EU) No. 540/2014 [17], which governs permissible noise levels from motor vehicles. The technical parameters for AVAS design and performance are based on the United Nations Economic Commission for Europe (UNECE) Regulation No. 138 [18], which has been transposed into EU law through the above regulatory instruments. Of note is that from 1 July 2021, the AVAS mandate was extended to all newly sold electric and hybrid vehicles within the EU, which was a transitional period intended to allow the vehicle industry to adapt to the new requirements. Since these regulations are not retroactive, many EVs manufactured before 2019 remain road-legal even though they lack any AVAS capability and are effectively silent at low speeds in built-up urban areas. In practice, this means that even vehicles produced as late as the first half of 2021 may still be on the road without AVAS installed, yet remain fully compliant with the law.

For those EVs that do come under these EU regulations, AVAS must activate automatically when the vehicle is moving at speeds below 20 km/h or when reversing. At speeds above 20 km/h, it is considered that the ambient noise produced by EVs is comparable to that of ICE vehicles, whereby the vast bulk of vehicular noise is dominated by tyre-on-road sound.

AVAS requirements currently apply to both vehicle categories M and N (passenger and goods transport vehicles), although increasing attention is also being directed toward smaller personal mobility devices (e.g., electric scooters), which pose similar risks but are typically regulated at municipal or national levels and lack a harmonized EU-wide framework.

To understand how AVAS requirements crystallized, it is useful to consider the minutes of UNECE's working groups - particularly the *Quiet Road Transport Vehicles* (QRTV) informal group (2010–2012) [19]. This group gathered inputs from national authorities, academics, industry, disability organizations, and trade representatives, balancing these perspectives and distilling a substantial body of third-party contributions. Responsibility then passed to the UNECE's *World Forum for Harmonization of Vehicle Regulations* (WP.29) [20], which oversaw the eventual publication of Regulation No. 138. Prominent inputs included surveys and market data from the Japan Automobile Standards Internationalization Center (JASIC); controlled trials by Germany's Bundesanstalt für Straßen- und Verkehrswesen (BASt); early U.S. research updates from the National Highway Traffic Safety Administration (NHTSA); and consolidated manufacturer positions via the Organisation Internationale des Constructeurs d'Automobiles (OICA). Importantly, these early discussions repeatedly highlighted perceptual concerns - such as detectability, locatability, attention-catching "departure" cues, and directivity - as well as scope debates (e.g., whether very quiet ICE vehicles or powered two-wheelers/e-bikes should be included).

As submissions progressed, however, some ambiguities also emerged in the way data were presented to the UNECE process. For instance, certain comparative trials reported EV detectability as equivalent to that of ICE vehicles, yet did not specify which ICE models were used as benchmarks. The absence of such detail makes it difficult to evaluate whether the chosen comparators may have been uncharacteristically quiet - as is the case with some luxury petrol vehicles - thereby potentially biasing the impression that EVs already achieved acceptable detectability levels as early as 2011. Such limited transparency may have reinforced a regulatory focus on minimal SPL thresholds, rather than encouraging a more comprehensive exploration of perceptual outcomes across varied real-world contexts. As a result, many of the recurring concerns raised during these formative discussions were only partially reflected in the final text of UNECE's 2016 regulation, which ultimately centered on minimum A-weighted sound levels, one-third-octave coverage, and a speed-correlated frequency shift - while leaving critical perceptual dimensions such as locatability and effective distance-estimation largely un-operationalized. Another illustration of certain compromises made during the regulatory process was the exclusion of Powered Two-Wheelers (PTWs), such as motorcycles, scooters, and micromobility devices from scope as early as 2011. This decision reflected what could be considered a commercially pragmatic narrowing of the regulatory frame-of-reference, prioritizing feasibility and industry consensus over the integration of dedicated, first-hand pedestrian studies in relation to PTWs.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Therefore, the UNECE - and by extension the European Union - did not build its regulation around a series of dedicated perceptual studies as did the NHTSA when devising U.S. AVAS regulation in line with the Pedestrian Safety Enhancement Act (PSEA) [21]. Nonetheless, the two regulatory pathways ultimately converged on broadly similar requirements, reflecting a shared emphasis on feasibility, measurability within existing acoustic test procedures, and commercial practicalities. While this alignment has ensured regulatory clarity and harmonization across jurisdictions, it has also meant that the richer potential of perceptual-based academic research has yet to be systematically integrated into either framework. Internationally, AVAS regulation has relatively few but sometimes notable differences [22]. In the United States, it is regulated by FMVSS No. 141 [23] and in China by GB/T 37153-2018 [24].

From a compliance-testing perspective, the methodologies used to assess AVAS performance are set out in ISO 16254 [25], a standard developed by the International Organization for Standardization (ISO). They provide a measurement framework rather than a pass/fail benchmark, as regulatory compliance is determined by whether a vehicle meets the thresholds specified in UNECE R138 and corresponding EU regulations, not by ISO 16254 itself. Instead, the ISO standard serves as the procedural basis for applying those criteria, ensuring harmonization across key acoustic properties such as sound pressure level (SPL), frequency content and tonality, as well as in the definition of pass-by measurement setups, instrumentation, and track/environmental specifications. Table I summarizes how ISO and UNECE complement one another in defining and enforcing AVAS compliance.

A. AVAS Specifications

The evolution of AVAS research reflects a trajectory from initial recognition of safety risks posed by silent vehicles [26][27][28] to the development of sophisticated sound-design models and compliance frameworks [29][30][31][32]. Indeed, this field is of considerable societal importance. As discussed in Section II, the rise in pedestrian accidents involving quiet vehicles has led to global legislation defining sound levels, spectral content, and activation rules, with the aim of improving detectability while limiting noise pollution [30][33][34]. Therefore, while manufacturers have flexibility to design their own sound signatures - which they often leverage for brand identity purposes - certain high-level criteria must be reached in accordance with the regulations. Table II outlines the key components of the AVAS regulation within the EU area.

B. Branding versus Function

Despite the presence of regulatory specifications, the real-world performance of AVAS continues to face growing scrutiny [10] [35][36]. Implementations differ markedly across manufacturers, and many AVAS designs fulfil a dual role: (1) providing a pedestrian safety warning and (2) contributing to the vehicle's sonic branding, whereby the sound symbolically conveys the identity or characteristics of the vehicle or manufacturer [37][38][39]. While existing regulatory frameworks seek to enhance detectability and limit annoyance [40][41][42], these goals may be compromised when AVAS signals are also shaped by branding requirements. In many cases, the resulting design choices are influenced more by aesthetic, marketing, or brand-identity considerations than by functional acoustic safety performance [43][44][45]. While aesthetic distinctiveness is recognized as a legitimate commercial objective, including within the regulatory framework, it cannot be allowed to undermine the functional performance required in safety-critical applications.

AVAS, therefore, offers an opportunity to strengthen interdisciplinary collaboration between creative professionals (e.g., composers, sound designers) and scientific experts in psychoacoustics, vehicle acoustics, and urban noise [34][46][47][48][49]. Such collaboration could ensure that aesthetic motivations do not inadvertently undermine AVAS' functional effectiveness in real-world pedestrian environments. Further examination may be needed to determine whether current regulatory frameworks and standardization processes adequately support such interdisciplinary engagement. In particular, it would be valuable to explore how recognizing the role of aesthetic branding could encourage greater inclusion of end-user perspectives, especially those of vision-impaired pedestrians.

As such, more inclusive, user-centered studies would provide critical empirical data to evaluate whether AVAS systems balance branding ambitions with perceptual clarity and functional reliability [50][51][52][53][54]. Indeed, although all AVAS signals in EU-category M and N vehicles must conform to specified thresholds and satisfy ISO-based compliance tests, the substantial design latitude afforded to sound designers has raised increasing concern over whether AVAS systems adequately meet the needs of pedestrians who rely on auditory cues [50][55]. Relatively few studies have engaged directly with vision-impaired pedestrians, leaving regulators and manufacturers with only a limited understanding of how these groups perceive and interpret different AVAS designs.

While UNECE R138 and ISO 16254 embed some psychoacoustic concepts (particularly tonality and modulation) into regulatory and testing frameworks [56]; they do so without explicitly and directly referencing some of the newer scientific literature that underpins those conventions. Therefore, it is perhaps legitimate to question how such regulatory requirements have been codified and subsequently operationalized, particularly when manufacturers face competing pressures between branding differentiation and compliance. Essentially, despite regulatory progress, a critical knowledge gap remains regarding the optimal balance between regulatory compliance, psychoacoustic effectiveness, and inclusive end-user acceptability [57][58][59][60].

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I. ISO 16254:2024 MEASUREMENT PROCEDURES AND UNECE REGULATION NO. 138 AVAS REQUIREMENTS. COMPARED TO ITS INITIAL 2016 EDITION, ISO 16254:2024 EXPANDS THE ACOUSTIC TESTING FRAMEWORK BY INCORPORATING ADDITIONAL MICROPHONES FOR PASS-BY DATA CAPTURE

Dimension	ISO 16254:2024 (Measurement Standard)	UNECE Regulation No. 138 (Regulatory Thresholds)
Purpose	Defines standardized methods for measuring AVAS sounds in controlled environments.	Sets mandatory performance thresholds and operational conditions for AVAS systems in electric and hybrid vehicles.
Scope	Applicable to M and N category vehicles at standstill and forward motion up to 20 km/h (also includes reverse).	Applies to M and N category vehicles; AVAS must operate up to 20 km/h and in reverse gear.
Minimum Sound Pressure Level	Measures A-weighted SPL[a] using 10-mic array (or full hemisphere in some setups); includes FFT[b] and psychoacoustic metrics.	Must emit at least 50 dB(A) at 0 km/h (stationary); at least 56 dB(A) at 10 km/h; at least 61 dB(A) at 20 km/h
Frequency Content	1/3-octave band spectrum[c] from 160 Hz to 5000 Hz is measured and reported	Must contain at least two 1/3-octave bands between 315 Hz and 5000 Hz, one of which must have the maximum SPL
Frequency Shift Requirement	Tonal shift (Δf) measured across speed; FFT used to evaluate frequency increase from 5 to 20 km/h	AVAS must show a frequency shift correlated with vehicle speed between 0 and 20 km/h; this can be achieved by modulation or increasing pitch
Tonality	Calculates tonal loudness, modulation, and tonality metrics (e.g., Zwicker method [56])	No numerical psychoacoustic threshold; requirement is that the sound must be recognizable and interpretable as vehicle motion by pedestrians (these are vague, whereas the updated ISO standard is now specific in what and how to measure)
Reverse Operation	Measures sound while reversing under standardized conditions	AVAS must emit sound when reverse gear is engaged, regardless of speed
Pause Function	Optional mute function noted in test description	Must not be continuously "pausable" by the driver; only temporary (e.g., for parking assist) interruptions are permitted
Microphone Configuration	10 microphones in linear or circular array (2 metres[d] from test line, 0.8–1.6 m height); also near-field mic for module testing, In the 2016 version of the standard, this comprised 2 microphones.	Test setup follows ISO 16254 (2016 version) - microphone position and track environment are indirectly inherited
Testing Environment	Outdoor (ISO 10844 compliant test track [61]) or indoor anechoic chamber; environmental conditions specified	Requires ISO 10844 test track or equivalent; temperature and wind speed constraints apply
Reporting Output	Full SPL and FFT data; metrics like $Lcrs_{10}$, Lst_{fwd}, Lst_{rev},[e] 1/3-octave spectrum plots	Pass/fail based on SPL thresholds, tonal content, and frequency shift behavior

a. Sound Pressure Level (SPL) is a ratio measure of the pressure variation in air caused by sound waves, relative to atmospheric pressure. When detected by a listener, the degree of this variation is perceived as loudness. Because SPL can be applied both in physical measurement and perceptual assessment, weighted scales are used to clarify what dimension is being addressed. When SPL is considered in relation to the minimum sound pressure required for human hearing, the A-weighted scale is typically used.
b. The Fast Fourier Transform (FFT) is a computational process that decomposes a captured and digitized acoustic signal into its constituent frequencies along the frequency domain. This decomposition reveals the individual frequency components that together form the complex waveform.
c. A 1/3-octave band spectrum is a way of grouping FFT information into frequency bands that are each one-third of an octave wide. An octave represents a doubling of frequency (for example, 500 Hz to 1,000 Hz) and dividing that range into three parts gives relatively narrow bands still useful for analysis.
d. The 2-meter measurement distance is a technical standard for repeatable testing, not a proxy for perceptual detection distance in real-world pedestrian contexts.
e. Shorthand for the measurement of SPL under certain vehicle conditions, where L=SPL; crs10 = constant running speed (10 km/h); stfwd = starting forward; strev = starting reverse.

TABLE II. IMPLEMENTATION OF UNECE INTO EU AVAS REGULATION

Action	Description
Automatic Activation	Must activate without driver intervention when the vehicle moves forward at speeds up to 20 km/h or when the vehicle is in reverse, regardless of speed.
Minimum Sound Pressure Levels (A-weighted SPL at 2 m)	50 dB(A) at 0 km/h (stationary, idling with AVAS on); 56 dB(A) at 10 km/h; and 61 dB(A) at 20 km/h
Frequency Content Requirements	Must include at least two 1/3-octave bands between 315 Hz and 5 kHz. One band must contain the maximum SPL while the second must be within 5 dB(A) of the maximum.
Speed-Dependent Frequency Shift	The signal must exhibit a perceptible increase in frequency (pitch or modulation) as the vehicle speed increases from 0 to 20 km/h. This shift must be continuous and clearly correlated with speed.
Reverse Operation Sound	When reversing, the AVAS signal must also operate and may be distinct from the forward sound. It must still meet minimum SPL and content requirements.
Tonal Quality and Recognisability	The sound must be recognizable as a vehicle in motion and it must be clearly audible in an urban background noise environment. No exact tonal or psychoacoustic metric is required in the current regulation, but tonality and modulation are implicitly expected to enhance detectability.
No Permanent Deactivation	A pause/mute function is not allowed unless it is temporary and is still required for specific operational scenarios (e.g., parking assist). The AVAS must resume automatically when the condition ends.
Compliance with Testing Protocols	Must be verifiable on an ISO 10844:2014 test track [61]. The new AVAS standard specifies additional requirements for outdoor test conditions, where no reflective surfaces can be within a 50m radius of the source and microphones. Measurement is performed with 10 microphones, placed at 2m distance and 0.8m-1.2 m height.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

III. METHODOLOGY

The EVA study employed a mixed-methods survey to investigate pedestrian experiences and perceptions of EVs and AVAS, with particular attention to individuals who rely on auditory cues for navigation. The survey was ethically approved by the Ethics Committee of the Technological University of the Shannon prior to distribution.

Participants were recruited internationally by reaching out to road safety organizations, visual-disability advocacy groups, and academic forums in auditory science. Dissemination channels also included targeted mailing lists, disability support networks, and social media platforms. Eligibility criteria required participants to be 18 years or older, capable of providing informed consent, and to have previously encountered one or more EVs (either visually or aurally) in real-world environments.

Accessibility was a core design principle of the survey instrument. The online survey platform (SurveyMonkey®) was tested and optimized for compatibility with screen readers and other assistive technologies, ensuring independent completion by individuals with vision-impairments. Vision-impairment status was self-reported by participants in line with National Disability Authority (NDA) definitions [62], encompassing both full and partial vision loss. In addition, to broaden accessibility and comprehension, the survey language was structured in accordance with guidelines from the National Adult Literacy Agency (NALA) [63] to ensure individuals with varying literacy levels across both groups could understand and engage with the questions effectively. To safeguard privacy, no personally-identifiable information was requested, and IP tracking was disabled. Informed consent was embedded within the survey introduction, and final submission was treated as confirmation of consent.

A total of 135 individuals responded to the survey, with initial data cleaning removing 25 respondents who disengaged after initial consent, leaving a total of 110 legitimate responses. Participants were asked to self-report sensory impairments, if any - see Table III. With Hearing Sensory Impairment (HIS) and Other Stated Impairment (OSI) groups having insufficient sample sizes for comparative statistical analysis, these respondents were subsequently excluded. The final cleaned dataset, therefore, comprised 90 individual responses: 51 participants self-reporting no sensory impairment (NSI) and 39 self-reporting a vision sensory impairment (VSI). Data were stored in CSV format and statistically analyzed using R Version 4.4.2.

TABLE III. RESPONDENT COUNT BROKEN INTO SENSORY CATEGORIES.

Category	Number
No Sensory Impairment (NSI)	51
Visual Sensory Impairment (VSI)	39
Hearing Sensory Impairment (HSI)	17
Other Stated Impairment (OSI)	3
TOTAL	**110**

A. NSI and VSI Groups

Figure 1 shows the age distribution of participants in NSI and VSI groups. While the VSI group had a slightly higher median age compared to the NSI group, the ranges overlap substantially. This therefore indicated that age was unlikely to be a confounding factor in interpreting the differences in experiences or perceptions between groups.

The survey itself comprised two components:

- A series of Likert-scale statements assessing perceptions of EV detectability and AVAS effectiveness, which progressed from an analysis of a smaller subset of early respondents published in Neff et al. [1].
- Open-ended questions inviting qualitative insights into pedestrian experiences, which also progressed from an analysis of a smaller subset of early respondents published in Neff et al. [64].

This necessitated a mixed-methods analysis combining descriptive statistics and comparative analysis, along with thematic coding and sentiment analysis to identify key perceptual differences between the cohorts. Note that response patterns varied across sections (see Table IV), with 48 NSI and 39 VSI participants completing the Likert items, and 39 respondents from each group providing qualitative responses.

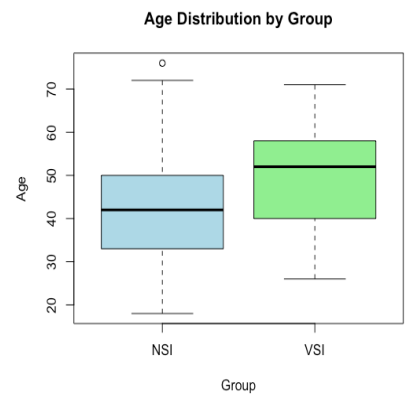


Figure 1. Age distribution of participants with NSI and VSI, shown as a boxplot. The VSI group shows a higher median age and narrower lower age range compared to the NSI group, which includes younger participants and one older outlier. The NSI Group's age range was 18 to 76 compared to the VSI Group's range of 26 to 71. The NSI Group's' age median was 42 compared to the VSI Group's median of 52; and the NSI Group's age mean was 42.44 compared to the VSI Group's mean being 49.06

TABLE IV. NUMBER OF PARTICIPANTS FOR EACH SURVEY SECTION BY GROUP.

Group	Total Participants	Likert Responses	Open-Ended Responses
NSI	51	48	39
VSI	39	39	39

B. Likert Statements

To capture quantitative perceptions of EV and AVAS experiences, the survey included a set of nine structured statements (see Table V) assessed using a 5-point Likert scale.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

These statements were designed to evaluate pedestrian views on EVs that they encountered and the detectability, interpretability, and overall effectiveness of AVAS. Participants were asked to indicate their level of agreement with each statement, reflecting their subjective experience and comfort in scenarios involving EVs.

Statements L1 through L8 used a common response scale as follows: 1 = I disagree a lot; 2 = I disagree just a little; 3 = I don't know; 4 = I agree just a little; 5 = I agree a lot.

Statement L9 maintained the same ordinal structure but was reworded to reflect perceived difficulty as follows: 1 = Very difficult; 2 = Difficult; 3 = Neither difficult nor easy; 4 = Easy; 5 = Very easy.

This approach allowed for the quantification of perceptual and behavioral dimensions of AVAS from both sighted and vision-impaired participants, enabling cross-group comparison. The inclusion of a neutral midpoint ("I don't know" or "Neither difficult nor easy") permitted non-binary responses and accommodated uncertainty where relevant.

TABLE V. THE NINE LIKERT STATEMENTS IN THE EVA SURVEY.

Code	Statement
L1	I feel safe when I think there might be an EV close by.
L2	It is easy to notice an EV approaching because of its sound.
L3	Sounds made by EVs help me understand what the vehicle is doing.
L4	I feel confident I understand an EV's next action based on its sound.
L5	I can react quickly to the sound of an EV when necessary.
L6	I find the sound of EVs pleasant.
L7	It takes little effort for me to listen to an EV's sound and understand what it is doing.
L8	I believe that the sound from all electric cars will be a positive thing for noise levels in busy cities and towns.
L9	Imagine you are standing on a busy street with lots of electric cars making sounds. Do you think it would be easy or hard to know when it is safe to cross the road?

C. Descriptive Analysis for Likert-Scale Items

Before conducting inferential statistical analysis, an initial descriptive analysis was performed to examine central response tendencies across the nine Likert-scale items. This was done to help determine if follow-on analysis was justified and also to contextualize overall response patterns that might highlight early group-level perceptual differences.

Given the ordinal nature of Likert data, the median serves as a more appropriate measure of central tendency than the mean. Therefore, each Likert item was calculated separately for the NSI and VSI groups, and responses were numerically coded on a 5-point ordinal scale, with consistent directional alignment across all statements (1 = most negative, 5 = most positive).

The median calculations were followed by a measure of the interquartile range (IQR), which provides an indication of the variability in participant responses within each group. Since normal distribution cannot be assumed in ordinal Likert responses, the IQR was deemed an appropriate non-parametric measure of statistical dispersion at the local group level. This served the purpose of giving the authors a vantagepoint of where the central 50% of responses fall (i.e., between the 25th and 75th percentiles), offering an indication of how consistent or varied participant ratings were for each statement. When interpreted alongside the median, the IQR allowed for a richer understanding of both central tendency and response spread.

In addition, effect size was calculated for each Likert item to estimate the magnitude of perceptual differences (if present) between the NSI and VSI groups. This was also done to estimated how substantial any inferential difference was in subsequent significance testing independent of sample size. Cliff's delta (δ) was selected as an appropriate non-parametric effect size measure, as again, it is suited to ordinal data. This statistical method quantifies the likelihood that a randomly selected participant from one group will have a higher or lower score than a participant from the other group. Values range from –1 to +1, where 0 indicates complete overlap, and values approaching –1 or +1 indicate stronger group-level divergence. When considered alongside medians and IQRs, Cliff's delta contributed to a more comprehensive understanding of group-level differences and highlighted which Likert items should undergo deeper inferential analysis.

D. Inferential Analysis for Likert-Scale Items

To move beyond descriptive summaries and examine whether observed group-level patterns reflect statistically meaningful differences, a series of inferential statistical analyses were conducted. These analyses aimed to explore both within-group associations and between-group differences in responses to the Likert items. This included both correlation analysis, significance testing, and multivariate analysis.

Correlation Analysis: To explore potential relationships between individual Likert-scale items within each participant group, Spearman's rank correlation coefficient was used. This non-parametric method assesses the strength and direction of monotonic associations between variables without assuming normal distribution. Separate correlation matrices were calculated for the NSI and VSI groups, allowing for the identification of within-group response patterns that may reflect underlying perceptual characteristics.

Significance Testing: To evaluate whether responses to Likert statements differed significantly between the NSI and VSI groups, Mann-Whitney U tests were performed for each Likert item. The Mann-Whitney U is a non-parametric test and suitable for independent samples and ordinal data, making it appropriate for comparing central tendencies between the NSI and VSI groups. In addition to p-values from the U tests, Rank-Biserial Correlation (r) was computed as an accompanying measure of effect size, offering insight into the magnitude and practical relevance of observed differences.

Multivariate Group Difference Analysis: To complement item-level testing and to examine overall response patterns across the full set of Likert items, a Permutational Multivariate Analysis of Variance (PERMANOVA) was conducted. This technique allowed us to compare group-level profiles across the multiple

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

dimensions using non-parametric permutation tests. A Gower distance matrix was used to account for the ordinal nature of our Likert responses. As only two groups (NSI and VSI) were compared, the test included just a single degree of freedom in the analysis.

Taken together, these inferential analyses were selected to provide a robust approach to analyzing our Likert data whilst being careful about assuming normal distribution. All in all, they helped us to understand both the structure and statistical significance embedded in the differences or similarities between the NSI and VSI groups when it came to documenting their experiences and perception of AVAS.

E. Thematic and Sentiment Analysis of Questions 1 and 2

In addition to the Likert-scale items, the survey included four open-ended prompts designed to elicit qualitative insights, including the way these responses were conveyed on a sentiment level. These qualitative questions were specifically developed to capture perceptual nuances, emotional responses, and real-world encounters that Likert-scale items often fail to reveal clearly. These questions and their associated research rationale are presented in Table VI.

In R, the *tidyverse* suite was used for data preparation, and *rstatix* supported nonparametric testing. Thematic coding was conducted using the *quanteda* package, while *sentimentr* was used to calculate polarity scores and examine emotional tone. Therefore, analysis of the open-ended questions was multi-phased, relying on a hybrid computational-linguistic workflow focusing initially on only Questions 1 and 2. Question 1 was treated as a potential source of thematically rich, content-driven information about participants' baseline mental models of EVs versus ICE vehicles, whilst Question 2 served to provide insight into the affective tone characteristics of EV acoustics.

The first phase of analysis involved dictionary-based thematic coding of Q1 using predefined keyword sets aligned to six categories – namely: Comfort; Audibility; Environment; Cost; Performance; and Safety. These themes were chosen to map both functional concerns (e.g., detectability or environmental benefit) and experiential attitudes (e.g., acoustic pleasantness or risk). Each response was parsed for the presence of these theme indicators, resulting in binary (TRUE/FALSE) flags for each participant across all six dimensions. This allowed direct comparison of theme frequency between NSI and VSI groups.

The thematic categories were derived through an iterative process informed by two sources: (1) a review of prior AVAS and transport acoustics literature, which frequently organizes perceptual responses along functional and experiential dimensions, and (2) scanning of a subset of responses to identify recurrent terms and concerns. This dual approach ensured that the coding framework captured both established research themes (e.g., detectability, environmental benefit) and emergent participant attitudes (e.g., acoustic pleasantness, perceived risk).

Following thematic tagging, group-level comparisons were conducted using frequency tables and Fisher's exact tests to assess whether theme prevalence differed significantly between NSI and VSI participants. This was supplemented by calculation-of-odds ratios and confidence intervals to support interpretability.

The second phase of analysis focused on responses to Q2 ("How would you describe the sound of an EV?"), which offered insight into participants' emotional and perceptual framing of EV acoustics. While Q1 explored broad conceptual differences between EVs and ICE vehicles, Q2 responses were more introspective and affective in nature. Each participant's Q2 response was analyzed to produce an average polarity score, ranging from –1 (strongly negative sentiment) to +1 (strongly positive sentiment). This continuous sentiment measure allowed for richer within-group variation to be captured, compared to simply binary-theme flags. Group-level summary statistics (i.e., mean, median, standard deviation, and proportions of negative, neutral, and positive responses) were calculated separately for NSI and VSI groups.

To assess whether the two groups differed in their sentiment distribution, a Wilcoxon rank-sum test was performed, which is a non-parametric alternative to the t-test more suitable for relatively small and potentially non-normally distributed samples such as in EVA. In addition, Cliff's delta was computed to quantify effect size and interpret practical significance. This combination of descriptive and inferential techniques allowed us to determine not just whether the groups differed, but also whether such differences had meaningful interpretive weight.

TABLE VI. OPEN-ENDED SURVEY QUESTIONS AND THEIR CORRESPONDING RESEARCH OBJECTIVES.

Code	Text	Research Purpose
Q1	How do you think EVs are different from other cars that use petrol or diesel?	To explore participants' baseline mental models of EV versus ICE vehicles. This provides insight into what characteristics (e.g., noise, environmental impact, performance) are most salient in shaping public attitudes.
Q2	How would you describe the sound of an EV?	To collect qualitative descriptors of EV acoustics, including specific features such as pitch, volume, and tonal quality. Responses provide insight into how these auditory features influence perceptions of comfort, safety, and detectability.
Q3	Have you ever had an experience where the sound of an electric car surprised you or caught you off guard?	To identify real-world examples of surprise or near-miss incidents, particularly those linked to detectability limitations. These responses offer a window into safety-related experiences and the emotional salience of such events.
Q4	What specific sounds or noises from electric cars do you find pleasant or unpleasant? Please describe them and explain why you feel that way.	To explore individual preferences and aversions related to EV sound characteristics. These insights help reveal what acoustic design elements may contribute to positive versus negative user experiences.

F. Analysis Methods for Open-Ended Questions 3 and 4

Questions 3 and 4 were designed to elicit more contextualized insights into participants' experiences and emotional reactions to EV sounds. To analyze these

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

responses, we adopted a mixed approach combining response rate comparisons, nonparametric inferential tests, and polarity-based sentiment scoring.

For Question 3, which asked participants whether they had ever been surprised or caught off guard by an EV, analysis began with a response rate comparison between the NSI and VSI groups. A binary indicator was created for each participant reflecting whether they provided a valid response to Q3. We used a Fisher's exact test to assess whether response rates differed significantly between groups, reflecting potential differences in lived experience or salience of EV-related incidents.

Next, a sentiment analysis was conducted on the submitted free-text narratives, again using the *sentimentr* package in R. Only those participants who submitted non-blank responses were included in this phase. These per-person polarity scores enabled group-level comparisons of the affective tone of reported experiences. We employed a Wilcoxon rank-sum test to consider group differences (if any) in sentiment scores, with Cliff's delta reported as an effect size estimate. This combined approach once again allowed us to examine both how often participants described surprise-by-silence events and how emotionally valenced those reports were.

For Question 4, which asked participants to describe specific EV sounds they found pleasant or unpleasant, we again began by comparing response rates between groups using a Fisher's exact test. This assessed whether VSI participants were more likely than NSI to volunteer their views on EV sound characteristics. As with Q3, we created a binary flag for each participant indicating whether a non-blank response was recorded.

We then applied sentiment analysis to the valid Q4 responses using the same *sentimentr* pipeline. Each participant's description of pleasant or unpleasant EV sounds was scored for polarity, enabling quantitative comparison of sentiment between the two groups. Again, we used the Wilcoxon rank-sum test to assess statistical differences, with Cliff's delta providing a nonparametric measure of effect size. While thematic coding was not applied to Q4 responses in this phase, qualitative inspection of the comments supported interpretation of whether perceived pleasantness was driven by aesthetic preferences or by more functional concerns such as detectability.

Together, these analytical steps for Q3 and Q4 aimed to reveal whether group-level differences in experience, concern, or emotional response were evident; not just in how often participants commented, but in how they characterized and emotionally evaluated those experiences. This allowed for a deeper understanding of how EV sound design intersects with real-world safety, accessibility, and affective perception, particularly for vision-impaired pedestrians.

IV. Results: Descriptive analysis of Likert items

This section presents the findings from qualitative analyses of the survey data. Results are organized thematically, beginning with the statistical comparison of Likert-scale items between sighted and vision-impaired participants, followed by multivariate analyses of response patterns. Section VII explores qualitative insights drawn from open-ended responses, providing a complementary perspective on participant experiences and perceptions of EV sounds.

A. Median Differences between NSI and VSI

Table VII presents the medians of both NSI and VSI groups across all nine Likert statements. From these results, vision-impaired participants have experienced EVs differently compared to sighted pedestrians.

TABLE VII. Median ratings for each Likert statement.

Statement (Condensed)	NSI	VSI
L1: I feel safe when I think there is an EV close by.	3	1
L2: It is easy to notice an EV approaching due to its sound.	2	1
L3: Sounds made by EVs help me understand what the vehicle is doing.	2	1
L4: I feel confident I understand an EV's next action based on its sound.	1	1
L5: I can react quickly to the sound of an EV when necessary.	3	1
L6: I find the sound of EVs pleasant.	4	2
L7: It takes little effort for me to listen to an EV's sound and understand what it is doing.	3	1
L8: I believe that the sound from all electric cars will be a positive thing for noise levels in busy cities and towns.	4	3
L9: Do you think it would be easy or hard to know when it is safe to cross the road in a busy EV environment?	3	2

The VSI group consistently reported lower median ratings across all nine Likert-scale items, suggesting a pronounced difference in how pedestrians with vision-impairments perceive and interpret the sounds of EVs compared to sighted participants. In terms of perceived safety (L1), VSI participants expressed markedly less confidence, with a median rating of 1, compared to a neutral 3 in the NSI group. Detectability of EVs by sound (L2, L3) also emerged as a key area of divergence, with VSI participants again rating 1, indicating significant challenges in perceiving or interpreting EV acoustics. This does not suggest that the VSI Group has any auditory deficit compared to the NSI Group, rather, the difference is more likely explained by the sensory context – as in, NSI participants can confirm the presence or absence of an EV through sight, whereas VSI participants must rely solely on auditory cues.

Both groups expressed low confidence in predicting vehicle intent (L4), with medians of 1, pointing potentially to a fundamental design flaw in AVAS signals, which may not communicate directional or behavioral intent effectively. Reaction time (L5) revealed another significant contrast, with NSI respondents rating 3 (neutral) and VSI rating 1, suggesting that participants with vision-impairments feel substantially less prepared to respond quickly to approaching EVs based on sound alone.

NSI participants found EV sounds more pleasant overall (L6) and were more optimistic about their impact on urban noise environments (L8). Conversely, VSI participants appeared more critical or uncertain in both respects, suggesting different acoustic and safety needs. Finally, in L9, which addressed the complexity of real-world street

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

navigation amid multiple EVs, VSI respondents gave a median of 2 (more difficult), while NSI participants remained neutral (3). This further reinforces that current AVAS implementations may not be meeting the navigational needs of vision-impaired pedestrians in dynamic environments.

B. Interquartile Ranges for NSI and VSI Groups

The IQR values reveal important differences in how consistently each group responded to the survey items. The VSI group showed lower variability across most items, with five of the nine Likert items showing an IQR of 0, indicating that at least half of the vision-impaired respondents gave the same or very similar responses (see Table VIII). This suggests high agreement or shared perception within this group, particularly in items L2, L3, L4 and L7; all of which deal with detectability and interpretability of EV sounds.

In contrast, the NSI group displayed greater dispersion in their responses, with IQRs ranging from 1 to 3. Notably, items such as L2 and L7 had IQRs of 3, reflecting substantial variability in how sighted participants perceive the usefulness and clarity of EV sounds. This may reflect the fact that sighted individuals rely less consistently on auditory cues, leading to broader variation in their experiences and interpretations. Interestingly, both groups shared a moderate degree of variability in response to the sound pleasantness item (L6) and urban noise attitudes (L8), suggesting more subjective or context-dependent opinions across cohorts on these dimensions. These IQRs highlight that whilst VSI participants tended to rate the AVAS signals more negatively, they also did so with greater internal consistency, indicating a shared perceptual experience.

TABLE VIII. INTERQUARTILE RANGE (IQR) FOR EACH LIKERT ITEM BY PARTICIPANT GROUP.

Code	NSI	VSI
L1	2.25	1
L2	3	0
L3	2	0
L4	2	0
L5	2	1
L6	2	2
L7	3	0
L8	2	3
L9	1	1

C. Overview of Group Response Density

To complement the median and IQR data, violin plots incorporating a kernel density estimate were generated to visualize the underlying shape and distribution of responses across both groups (see Figure 2). These plots highlight not only central tendencies but also the frequency and symmetry of responses, revealing several notable patterns across Likert statements for each group. Across most items (L1 to L8), the NSI group demonstrated a broader spread of responses and more positively skewed distributions than the VSI group. For instance, statements L1, L2, and L5 show a visibly higher concentration of NSI responses in the upper end of the scale, with little to no overlap with VSI distributions, suggesting strong perceptual divergence. In contrast, items like L6 and L8 show slightly more overlap, although a positive bias remains evident in the NSI group.

Particularly notable are the relatively narrow and consistent response patterns within the VSI group across statements L2 to L4 and L7, where responses cluster tightly at the lowest scale values. This reinforces earlier descriptive findings (medians and IQRs) and suggests a high degree of consensus in negative perceptions within the VSI cohort for these statements.

D. Estimating Magnitude of Difference

While distribution plots and descriptive measures such as medians and IQRs offer insight into general response patterns, they do not quantify per se the magnitude of difference between groups. To address this, effect size statistics (δ) were calculated for each Likert item to estimate the strength of association between group membership (NSI vs VSI) and response level.

Importantly, effect size also serves a complementary role to significance testing. While p-values test whether an observed difference is likely due to chance, they are sensitive to sample size and do not convey the magnitude of that difference. However, a large effect size accompanied by a statistically significant p-value provides strong evidence that the group difference is both reliable and meaningful in practical terms.

The effect size results (see Table IX) reveal substantial group-level divergence between NSI and VSI participants in their responses to almost all survey Likert items. Notably, seven out of nine statements yielded large effect sizes ($|\delta| > 0.474$), indicating that participants in the VSI group were consistently more likely to rate items negatively and the NSI group more positively; a pattern that points to systematic differences in experience with EVs.

L1 and L2, which focus on the general detectability and clarity of EV detection, show the largest effect sizes ($|\delta| = 0.619$ and 0.617, respectively). This suggests that NSI participants may be more confident or comfortable with EVs in daily contexts, potentially because NSI pedestrians rely less on auditory cues for safety and spatial awareness.

L5 to L8 also register large effect sizes, reflecting perceptual dimensions such as ease of localization, emotional comfort, and confidence when crossing roads near the presence of EVs. This suggests that the auditory features of EVs may be less salient than intended, insufficient on their own to enable consistent detection, and ultimately require visual confirmation to ensure safe navigation. This ambiguity may play a role in VSI pedestrians' degree of trust (or lack thereof) in the acoustic cues embodied in AVAS.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

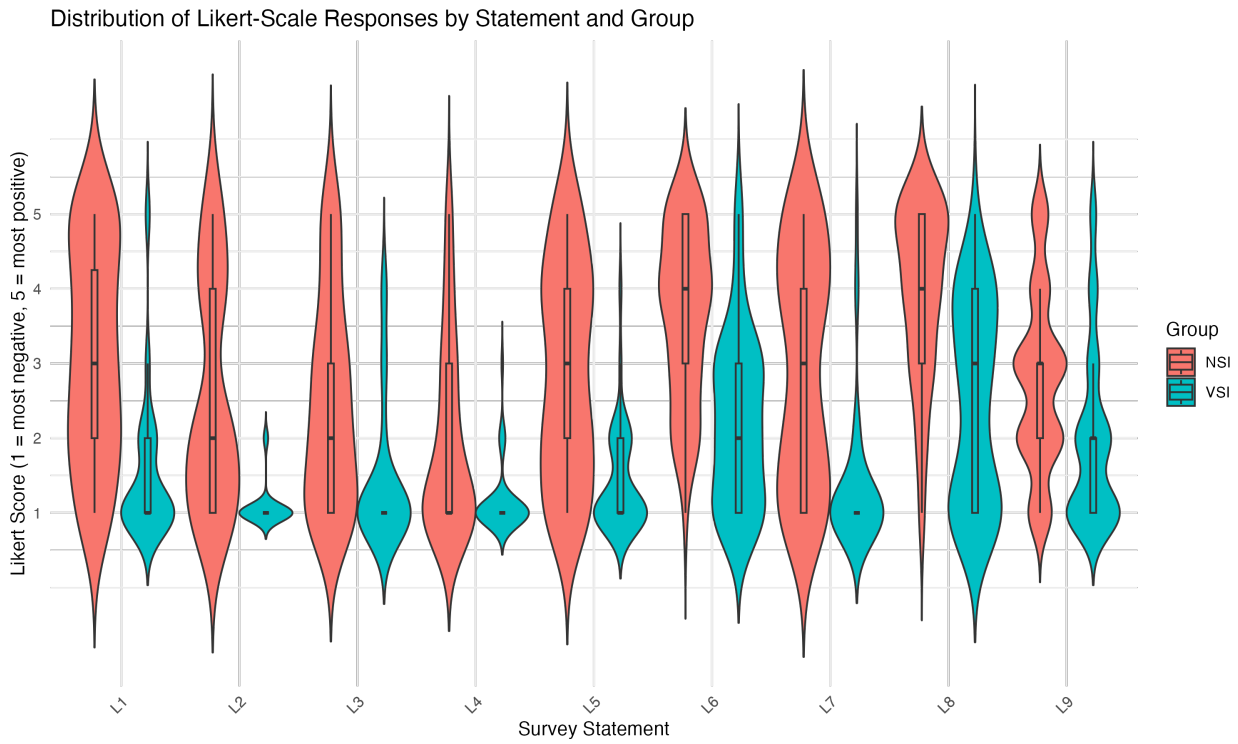


Figure 2. Violin plots of Likert-scale responses across nine statements, grouped by NSI (red) and VSI (blue). Each plot shows response density, with embedded boxplots indicating median and interquartile range (IQR). Higher Y-axis values mean more positive responses, broader X-axis ranges mean more within-group consensus.

TABLE IX. CLIFF'S DELTA (δ) EFFECT SIZE ESTIMATES FOR EACH LIKERT-SCALE ITEM COMPARING NSI AND VSI GROUPS.

Code	Delta	Magnitude
L1	0.619	Large
L2	0.617	Large
L3	0.46	Medium
L4	0.36	Medium
L5	0.589	Large
L6	0.602	Large
L7	0.589	Large
L8	0.584	Large
L9	0.404	Medium

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

The medium effect sizes observed in L3, L4, and L9 suggest more modest but still meaningful divergences. For example, L3 and L4, which capture aspects of perceived risk or uncertainty, may reflect more subjective or situational variability, even among VSI participants. The medium effect in L9, which concerns overall perceived safety in near-future environments occupied by many EVs, points to a nuanced area where both groups share some common ground.

Taken together, these effect-size magnitudes suggest that the acoustic characteristics and perceived behavioral implications of EV sounds are experienced quite differently depending on if a pedestrian has a vision sensory impairment or not. The consistently high δ values in favor of the NSI group suggest that VSI participants are not receiving clear or unambiguous auditory cues from the AVAS systems used in EVs, highlighting a potential gap in the effectiveness of these signals for vision-impaired pedestrians.

V. Results: Correlation Analysis on Likert Items

Spearman's rank correlation was employed to quantify the strength and direction of monotonic associations between pairs of Likert-scale statements. This non-parametric method allowed us to explore whether higher agreement with one statement corresponded with higher (or lower) agreement with another. Separate correlation matrices were generated for the NSI and VSI groups to examine how patterns of internal consistency, alignment, or divergence differed within each participant cohort. Table X presents Spearman's ρ (rho) correlation coefficients for the NSI group, while Table XI displays the corresponding values for the VSI group.

Figure 3 visualizes the Spearman's rank correlation coefficients between all pairs of Likert items for the NSI group using a heatmap. The darker blue shades represent stronger positive correlations, while lighter shades indicate weaker associations. This matrix reveals several strong and consistent correlations among items related to auditory awareness, detectability, and clarity (e.g., L1–L5, L1–L6, L2–L3), suggesting that participants without sensory impairments formed coherent impressions of the AVAS stimuli. The strong correlations indicate that when one aspect of AVAS performance (e.g., sound detection) was rated highly, other related aspects (e.g., sound clarity or usefulness) were also likely to receive high ratings. Conversely, minimal correlations involving L7–L9 suggest more varied responses to statements relating to real-world decision-making or ambiguity, indicating greater diversity of interpretation in those areas.

Figure 4 presents the Spearman's rank correlation matrix for the VSI group. Compared to the NSI group, stronger correlations are concentrated around statements L3 to L5, indicating that participants with vision-impairments tended to show higher internal consistency when rating items related to clarity, usefulness, and environmental context. The very strong relationship between L3 and L4 suggests a shared perceptual framing of these statements; likely rooted in lived-experience that calls for navigating environments without visual information. Correlations across other item pairs are generally lower or more varied, reflecting greater diversity in how individual VSI participants interpreted or experienced AVAS cues across dimensions such as detection (L1) and confidence (L9). This pattern may point to variability in how effective current AVAS implementations are at delivering consistent or meaningful messaging for vision-impaired users.

TABLE X. Spearman's ρ correlation matrix for the NSI group, showing strength and direction of relationships between responses to different Likert-scale statements

	L1	L2	L3	L4	L5	L6	L7	L8	L9
L1	1.00	0.53	0.47	0.53	0.58	0.62	0.07	0.29	0.51
L2	0.53	1.00	0.55	0.43	0.52	0.27	0.11	0.14	0.36
L3	0.47	0.55	1.00	0.54	0.53	0.23	0.17	0.38	0.37
L4	0.53	0.43	0.54	1.00	0.53	0.30	0.20	0.24	0.27
L5	0.58	0.52	0.53	0.53	1.00	0.31	0.22	0.18	0.49
L6	0.62	0.27	0.23	0.30	0.31	1.00	0.17	0.43	0.20
L7	0.07	0.11	0.17	0.20	0.22	0.17	1.00	0.03	- 0.01
L8	0.29	0.14	0.38	0.24	0.18	0.43	0.03	1.00	0.22
L9	0.51	0.36	0.37	0.27	0.49	0.20	- 0.01	0.22	1.00

TABLE XI. Spearman's ρ correlation matrix for the VSI group, indicating the strength and direction of relationships between responses to each Likert-scale statement.

	L1	L2	L3	L4	L5	L6	L7	L8	L9
L1	1.00	0.16	0.21	0.31	0.35	0.17	0.27	0.13	-0.03
L2	0.16	1.00	0.32	0.45	0.37	0.16	0.10	0.26	-0.01
L3	0.21	0.32	1.00	0.74	0.43	0.49	0.34	0.31	0.40
L4	0.31	0.45	0.74	1.00	0.55	0.32	0.38	0.34	0.30
L5	0.35	0.37	0.43	0.55	1.00	0.24	0.31	0.41	0.33
L6	0.17	0.16	0.49	0.32	0.24	1.00	0.19	0.46	0.28
L7	0.27	0.10	0.34	0.38	0.31	0.19	1.00	0.39	0.32
L8	0.13	0.26	0.31	0.34	0.41	0.46	0.39	1.00	0.22
L9	-0.03	-0.01	0.40	0.30	0.33	0.28	0.32	0.22	1.00

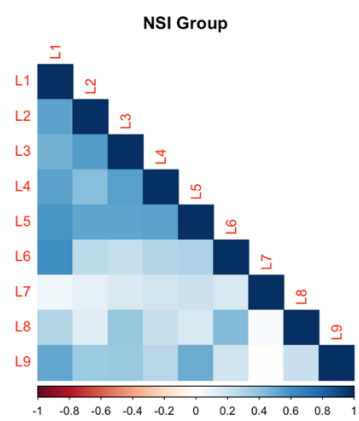


Figure 3. Spearman's rank correlation heatmap for the NSI group. Each cell shows the strength of monotonic association between pairs of Likert-scale statements, with color intensity representing correlation strength (ρ). Stronger positive correlations are indicated by darker blue. The pattern suggests coherent perceptions of AVAS features among NSI participants, particularly across L1–L6.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

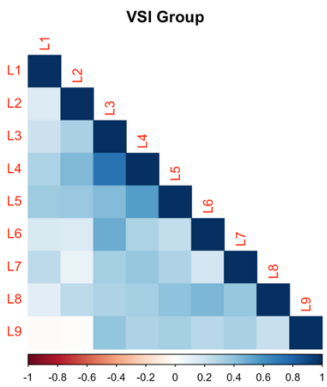


Figure 4. Spearman's rank correlation heatmap for the VSI group. Strong positive correlations are most prominent between Likert items L3–L5, suggesting coherent responses around statements relating to AVAS clarity and usefulness. Broader variation in correlation strength indicates more heterogeneous interpretation across other perceptual dimensions compared to the NSI group.

VI. Results: Significance Testing on Likert Items

To assess whether the NSI and VSI groups differed significantly in their responses to each AVAS-related Likert statement, a series of Mann-Whitney U tests were conducted. For each item, the U statistic and p-value were computed to determine the statistical significance of group differences. In addition, the Rank-Biserial Correlation (r) was calculated to provide a measure of effect size, offering insight into the practical magnitude of any observed differences. The results are presented in Table XII, highlighting both statistical and practical significance across the full set of statements.

This was followed by a multivariate analysis (PERMANOVA) to determine whether the overall response patterns across all Likert statements differed significantly between NSI and VSI groups. A Gower distance matrix was employed with 1 degree of freedom (df) as group difference was the single dimension under consideration. Spearman's rank correlation was employed to quantify the strength and direction of monotonic associations between pairs of Likert-scale statements.

A. Group Differences in Likert Responses

Table XII presents the results of Mann-Whitney U tests (U), p-values, Z scores, and rank-biserial correlation coefficients (r) as measures of effect size. All nine Likert items yielded statistically significant differences ($p < 0.001$) between the NSI and VSI groups. Of these, six items L1 (detectability of EVs by sound); L2 (clarity of EV approach); L5 (sound helped judge vehicle location); L6 (sound helped judge vehicle movement); L7 (sound gave enough warning); and L8 (sound helped feel safe) showed large effect sizes ($r \geq 0.5$), indicating substantial differences in perceptual experience between the groups. Items L3 (awareness of vehicle presence), L4 (ability to detect direction), and L9 (need for visual confirmation) demonstrated medium effect sizes, suggesting meaningful but more moderate group differences.

The strongest effects were observed for L2 ($r = 0.599$) and L1 ($r = 0.552$), highlighting that VSI participants were significantly less likely than their NSI counterparts to find the sound of EVs sufficiently clear or noticeable. This most likely reflects not a sensory deficit on the part of VSI respondents, but rather that NSI participants can supplement or bypass auditory ambiguity through visual verification, whereas those with vision-impairments are more vulnerable to the limitations in AVAS design.

Similarly, consistent large effects across L5 to L8 suggest that NSI pedestrians may not depend on AVAS to convey spatial or safety-related cues, while VSI participants (who must rely on these sounds) perceive them as falling short of their intended function. In this sense, the findings do not imply that VSI respondents struggle more with sound itself, but rather that the current auditory cues are not fit for purpose when visual compensation is not possible. These findings point to a systematic divergence in experience between groups, which is while AVAS may be minimally functional for those with full vision, it does not adequately support the needs of vision-impaired pedestrians. The consistently large and medium effect sizes, coupled with highly significant p-values, confirm that these differences are both statistically robust and practically meaningful.

TABLE XII. Results of Mann–Whitney U tests comparing responses to each Likert statement between NSI and VSI groups.

Statement	U	p-value	Z	r	Magnitude
L1	1516	<0.001	5.15	0.552	Large
L2	1514	<0.001	5.59	0.599	Large
L3	1366	<0.001	4.02	0.431	Medium
L4	1272	<0.001	3.51	0.376	Medium
L5	1487	<0.001	4.95	0.530	Large
L6	1500	<0.001	4.91	0.526	Large
L7	1487	<0.001	5.09	0.546	Large
L8	1482	<0.001	4.78	0.513	Large
L9	1314	<0.001	3.33	0.357	Medium

B. Group Difference in Overall Response Pattern

To assess whether the overall patterns of Likert-scale responses differed significantly between groups, a Permutational Multivariate Analysis of Variance (PERMANOVA) was conducted using a Gower distance matrix and 999 permutations. This non-parametric analysis tests whether overall response patterns across the nine statements differ by group (NSI vs. VSI). As shown in Table XIII, the PERMANOVA revealed a statistically significant group difference ($p = 0.001$), with the grouping factor (NSI vs. VSI) explaining approximately 30.6% of the variance in response patterns ($R^2 = 0.30583$). This indicates a meaningful divergence in how the two groups rated the statements when considered collectively, not just item by item.

A Principal Coordinates Analysis (PCoA) visualization (see Figure 5) based on the Gower distance matrix further illustrates this divergence, with a clear separation between NSI and VSI participants in the multivariate response space. The elliptical contours around each group represent 95% confidence intervals, reinforcing the systematic nature of this separation.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

These findings therefore reinforce the earlier Mann–Whitney U test results by showing that the divergence between NSI and VSI groups is not only significant at the individual Likert-statement level, but also consistent and substantial when evaluating the overall configuration of their responses.

A remaining 69.4% of variance remains unexplained and this residual variation may reflect individual differences unrelated to sensory status - such as prior exposure to EVs, age, mobility habits, mobility routes, cognitive strategies, or situational context during sound encounters (e.g., background noise, urban design, time of day). These factors, while outside the scope of the present analysis, merit future exploration to further characterize how pedestrians interpret AVAS cues.

TABLE XIII. PERMANOVA RESULTS COMPARING OVERALL RESPONSE PATTERNS BETWEEN NSI AND VSI GROUPS. PERMANOVA BASED ON GOWER DISTANCE WITH 999 PERMUTATIONS.

Source	df	Sum of Squares	R^2	F	p-value
Group	1	2.0421	0.30583	37.448	<0.001
Residual	85	4.6353	0.69417		
Total	86	6.6774	1.00000		

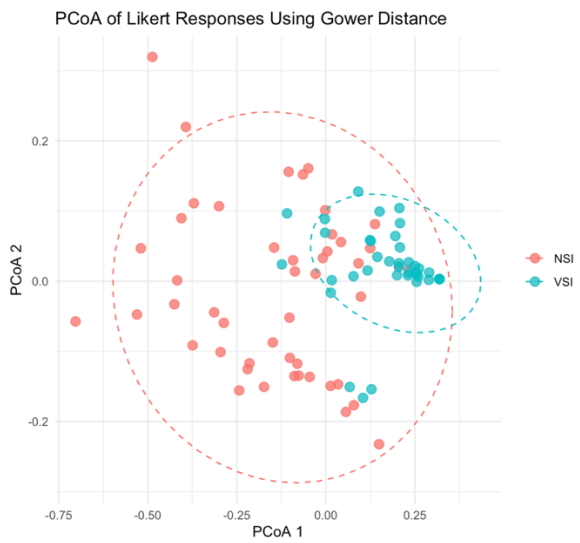


Figure 5. PCoA plot based on Gower distances calculated from Likert-scale responses (L1–L9). Each point represents an individual respondent, coloured by group. Ellipses indicate 95% confidence intervals for each group. The clear separation of the VSI cluster supports the significant group difference detected in the PERMANOVA ($R^2 = 0.306$, $p < .001$).

VII. RESULTS: ANALYSIS OF OPEN-ENDED RESPONSES

A. Analysis of Question 1

Thematic coding of Question 1 responses revealed that *Audibility* was the only theme to exhibit a statistically significant difference in prevalence between NSI and VSI participants. The odds of *Audibility*-related references were over ten times higher among one group compared to the other (OR = 10.21, 95% CI [1.25, 475.92], $p = 0.013$). This marked effect contrasts with the remaining five themes, for which no statistically reliable group differences were observed. Although *Comfort* and *Environment* featured moderately in both groups, the associated odds ratios hovered close to unity and confidence intervals were wide, indicating negligible and uncertain group-level variation. *Cost* and *Performance* were comparatively infrequent, and *Safety* was rarely mentioned at all, producing unstable estimates with correspondingly broad confidence bounds.

When considered together, these results suggest that perceptions linked to *Audibility* form the primary thematic distinction between NSI and VSI respondents when articulating how EVs differ from ICE vehicles (Q1), whereas other evaluative dimensions appear more uniformly distributed across groups (see Figure 6).

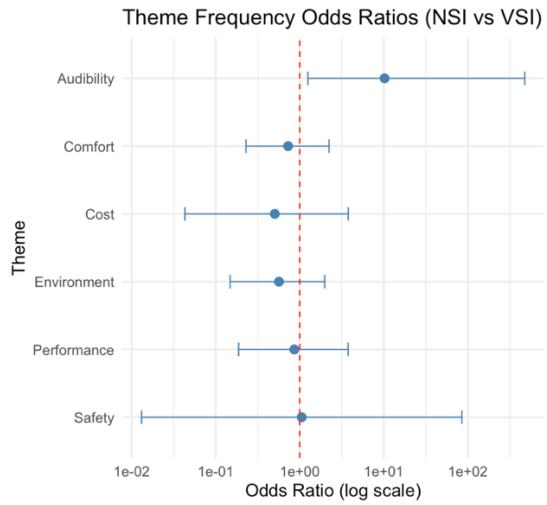


Figure 6. Theme frequency odds ratios (NSI vs VSI) for Q1. The plot shows odds ratios with 95% confidence intervals on a log scale.

B. Analysis of Question 2

A polarity-based sentiment analysis was conducted on participants' open-text responses to "How would you describe the sound of an EV?" This approach assigns each response a sentiment score ranging from –1 (entirely negative tone) to +1 (entirely positive tone), with scores near zero indicating a neutral tone. The analysis found that mean sentiment was positive for both groups, with NSI respondents showing a slightly higher average sentiment (M = 0.24, SD = 0.68) compared to VSI respondents (M = 0.16, SD = 0.64). These values suggest that, on average, both groups tended to use language with a mildly positive valence when describing EV sounds themselves, although variability within each group was also high.

C. Analysis of Question 3

Sentiment analysis revealed notable differences in the emotional tone of surprise experiences between groups (see Table XIV).

VSI participants demonstrated greater emotional variability in their surprise experiences (SD = 0.216 vs 0.121) and were three times more likely to report negative sentiment (31.6% vs 10.3% negative responses). While both groups

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

clustered around neutral sentiment overall, the broader range and higher frequency of negative responses among VSI participants suggests these surprise-encounters represent genuine safety concerns rather than neutral observations. The polarized nature of VSI responses (ranging from highly negative (-0.288) to highly positive (0.8)), indicates more emotionally charged experiences, consistent with the safety implications of reduced EV detectability for vision-impaired pedestrians.

However, while the descriptive statistics above suggest VSI participants tend to experience more negative sentiment, the formal statistical test fails to confirm this pattern (see Figure 7 where $p = 0.104$). In addition, the small effect size ($\delta = 0.215$) should also be interpreted cautiously, as it may reflect sampling variability rather than a true group difference. With 77 total responses, the study had a reasonable sample size relative to how many people tend to respond to such niche surveys, so there was a relatively good chance that meaningful differences would be detected in this question if they existed. The non-significant result therefore suggests that both groups experience similar emotional responses to EV surprise encounters, regardless of sensory status; and indicates that while VSI participants may face greater practical safety challenges with EV detection, their emotional reactions to surprise incidents are not systematically different from those of NSI participants. Note that this similarity in emotional impact likely does not validate current AVAS design but rather indicates that inadequate acoustic warning systems may be equally problematic for all pedestrians, eliciting comparable levels of distress across sensory abilities. Further research may be warranted, therefore, to develop acoustic cues that provide an advance "heads-up" alert on approach, helping to reduce the anxiety or discomfort associated with such encounters.

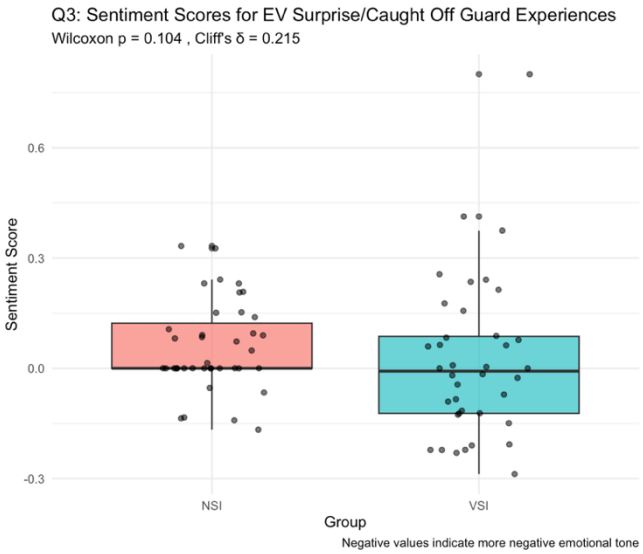


Figure 7. Sentiment scores for Q3 responses describing experiences of being surprised or caught off guard by electric vehicles, by participant group. Box plots show median, quartiles, and range; individual data points are jittered to show distribution.

D. Analysis of Question 4

An examination of Q4 revealed distinct patterns in the underlying motivations driving participants' evaluations of EV sounds (see Table XV), demonstrating a clear reversal in priorities between groups. NSI participants predominantly framed their responses around aesthetic preferences (50.0%), focusing on whether EV sounds were pleasant, soothing, or enjoyable. In contrast, VSI participants primarily emphasized functional concerns (51.5%), concentrating on detectability, warning capability, and safety implications. This pattern suggests that when evaluating EV acoustic characteristics, NSI participants approach the question from a comfort and pleasantness perspective, while VSI participants evaluate the same sounds through a safety and utility lens. Also, the higher proportion of mixed/unclear responses among VSI participants (33.3% vs 23.3% for NSI) may reflect the complex interplay between aesthetic judgment and practical necessity when auditory information is critical for navigation.

Indeed, follow-on sentiment analysis of Q4 responses revealed a statistically significant difference between groups when participants evaluated pleasant and unpleasant characteristics of EV sounds (Figure 8). NSI participants demonstrated significantly more positive sentiment (Wilcoxon $p = 0.024$, Cliff's $\delta = 0.331$, medium effect size) when describing EV acoustic features compared to VSI participants. This finding contrasts with the Q3 results, where no group differences emerged in emotional responses to surprise encounters. The significant difference in Q4 suggests that when explicitly evaluating EV sound characteristics, participants' sensory status may fundamentally influence their assessment. NSI participants, who predominantly focus on aesthetic qualities, tend to frame EV sounds more positively, likely emphasizing features such as quietness as pleasant or soothing. Conversely, VSI participants, whose evaluations center on functional concerns, express more negative sentiment, reflecting frustration with inadequate detectability.

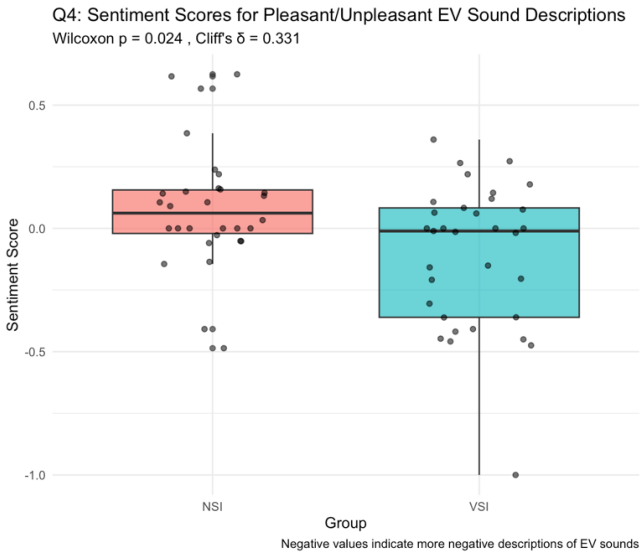


Figure 8. Sentiment scores for Q4 responses describing pleasant and unpleasant characteristics of EV sounds, by participant group.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE XIV. PERMANOVA RESULTS COMPARING OVERALL RESPONSE PATTERNS BETWEEN NSI AND VSI GROUPS. PERMANOVA BASED ON GOWER DISTANCE WITH 999 PERMUTATIONS.

n	mean_sentiment	median_sentiment	sd_sentiment	min_sentiment	max_sentiment	negative_responses	neutral_responses
NSI (39)	0.05657037	0	0.12122559	-0.1668115	0.3328201	4	24
VSI (38)	0.01913296	-0.0078763	0.21590186	-0.2877128	0.8	12	17
				positive_responses	**%_negative**	**%_neutral**	**%_positive**
			NSI (39)	11	10.3	61.5	28.2
			VSI (38)	9	31.6	44.7	23.7

TABLE XV. PERMANOVA RESULTS COMPARING OVERALL RESPONSE PATTERNS BETWEEN NSI AND VSI GROUPS. PERMANOVA BASED ON GOWER DISTANCE WITH 999 PERMUTATIONS.

Group	Primary Concern	n	%
NSI	Aesthetic	15	50
NSI	Functional	8	26.7
NSI	Mixed/Unclear	7	23.3
VSI	Aesthetic	5	15.2
VSI	Functional	17	51.5
VSI	Mixed/Unclear	11	33.3

E. Summary of Open-Ended Results

The analysis of open-ended responses reveals an interesting progression of how sensory status may shape EV-related perceptions and experiences. VSI participants demonstrated fundamentally different baseline mental models of EVs, being ten times more likely to conceptualize them through an audibility framework (Q1), establishing distinct cognitive foundations between groups starting off. While both groups initially described EV sounds with similar mildly positive sentiment (Q2), their evaluative frameworks diverged sharply when explicitly assessing pleasant/unpleasant sound characteristics (Q4); with NSI participants focusing on aesthetic qualities and expressing more positive sentiment versus VSI participants emphasizing functional concerns and expressing significantly more negative evaluations.

Interestingly, both groups demonstrated comparable emotional distress when describing surprise encounters with EVs (Q3), suggesting that there may be AVAS inadequacies that create universal pedestrian distress in surprise scenarios rather than this being an VSI-specific issue. This pattern indicates that EVs fail both groups but for different reasons: VSI participants experience inadequate functional warning (possibly being distance independent), while NSI participants still encounter distressing surprise incidents that suggest insufficient acoustic range or volume to provide adequate advance warning at appropriate distances.

Collectively, these findings suggest that acoustic stimuli may be processed through fundamentally different perceptual and evaluative frameworks depending on sensory status, yet result in similar emotional distress across groups. This indicates that there may still be a need to further refine AVAS design for addressing both functional detectability requirements and adequate detection distances for universal pedestrian comfort. This challenges the assumption that current acoustic warning systems adequately serve any pedestrian population through either sound quality or propagation characteristics.

VIII. DISCUSSION

A. Notification Timing, Distance, and AVAS Design

The findings of the EVA study highlight a systematic divergence in how sighted and vision-impaired pedestrians perceive and respond to EV sounds and their associated AVAS. Across all nine Likert-scale statements, participants with vision-impairments consistently reported more negative perceptions of safety, detectability, and interpretability compared to sighted participants. This difference was not only statistically significant but also accompanied by large effect sizes in many items, indicating practical, real-world implications for pedestrian safety.

A key insight emerging from the study is the role of distance and notification timing in shaping pedestrian experience. Participants, particularly those with vision-impairments, reported difficulties in detecting EVs early enough to allow safe and confident reactions. Surprise encounters were common across both groups, suggesting that the current design of AVAS may not provide adequate advance warning on a universal level. While regulations specify minimum sound levels and frequency characteristics, they do not explicitly address how far in advance an EV should be detectable or whether pedestrians should be given a clear “heads-up” alert to reduce surprise. In essence, distance is operationalized only in terms of measurement geometry and

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

not in terms of real-world perceptual outcomes. This regulatory vagueness leaves a critical gap: where compliance testing ensures only that vehicles emit sound at measurable levels but has no guidance on how that sound should function with distance in real urban environments to ensure meaningful detectability for all users.

The EVA findings on late or surprise detection of EVs resonate strongly with concerns already documented during UNECE's early QRTV sessions. For example, German BASt trials in 2011 [19] demonstrated that some ICE comparators - even those used in other submitted trials as benchmarks - were effectively as quiet as EVs at low speeds, raising questions about whether "quietness" should be treated as an EV-only issue. EVA participant-reports of delayed detection and reliance on visual cues suggest that this ambiguity persists in real-world pedestrian experience, and that the current minimum thresholds specified in UNECE R138 do not sufficiently account for this overlap. Similarly, EVA survey responses describing unexpected encounters with smaller micromobility devices echo an omission in the regulations: despite early debate during 2010-2011 about whether e-bikes or powered two-wheelers should be included, the scope was ultimately restricted to M and N categories (passenger cars and light goods vehicles). Considered together, the EVA results confirm that unresolved issues identified in UNECE's formative years remain active sources of risk and uncertainty in today's pedestrian experience more than a decade later.

Indeed, sounds may meet regulatory thresholds at close range on controlled test-tracks, but may fail to propagate effectively in dynamic, noisy environments where early detection is essential. Recent experimental work supports this distinction between detectability and functional perceptual efficiency. For example, Müller, Forssén, and Kropp [65] found that although AVAS signals can be detectable in controlled conditions, localization accuracy deteriorates significantly, particularly for tonal signals and worsens further when multiple vehicles are present. This highlights a limitation in standards whereby they equate threshold detectability with effective pedestrian awareness.

Moreover, while the updated ISO 16254 standard introduces more sophisticated measurement techniques with expanded microphone arrays and more refined psychoacoustic metrics, it still frames AVAS evaluation in terms of signal production and measurement, rather than alert effectiveness as a function of distance, pedestrian experience, or real-world noisy environments. Many academic studies have largely mirrored this focus, emphasizing laboratory-based detection thresholds and signal characteristics, but relatively little work has been done on how AVAS design could be optimized to improve anticipatory awareness and reduce surprise under real-world conditions. The forthcoming amendments in R138.02, which explicitly integrate these ISO updates into the regulatory framework, reflect this same trajectory: the regulation evolves by refining *how* signals are measured (e.g., multi-microphone arrays, clearer specifications for reverse sound) without substantively expanding *what* is measured. This iterative loop between UNECE regulation and ISO procedure ensures technical consistency but risks reinforcing a narrow emphasis on compliance metrics, rather than addressing the broader perceptual challenges consistently highlighted in end-user studies such as EVA.

B. Functional Meaning versus Branded AVAS

Another critical insight from the EVA study concerns the potential lack of functional meaning in current AVAS implementations. While EVs are required to emit sound that varies with speed, the standards require relatively coarse tonal features, leaving substantial latitude for car manufacturers to design uniquely branded sound-signatures. This flexibility has led to the emergence of distinctive acoustic identities for EVs, often emphasizing marketing differentiation over communicative function. Although such sounds can pass regulatory compliance tests, they may be failing to provide pedestrians with actionable cues about vehicle state or even driver intent as ICE vehicles often inherently do.

The survey results suggest that vision-impaired participants in particular seek sounds that are informative rather than aesthetic. Their framing of EV sounds as tools for detection and orientation contrasts with the sighted participants' emphasis on sound quality or pleasantness. This divergence highlights a tension in AVAS design: while branding may enhance product identity and user-experience for drivers, it risks undermining the functional purpose of AVAS for pedestrians who depend on it.

In terms of functional cues, ICE vehicles provide a richer palette of acoustic information than simply indicating speed, reverse, or stationary statuses. While these sounds may not always be aesthetically pleasing, they are closely tied to the mechanical operations of the engine. Subtle variations in engine tone can signal acceleration or deceleration, turning intent, driver impatience, even driver experience-level, as well as conditions such as engine misalignment on loose terrain or maneuvering in constrained spaces. Such cues give pedestrians valuable anticipatory information, helping them navigate traffic environments more safely and confidently. By contrast, current AVAS implementations often reduce this communicative richness to uniform, synthetic tones that satisfy baseline detectability requirements but lack the dynamic subtleties needed to support true situational awareness.

Future research should therefore extend beyond compliance-based assessments to examine how ICE soundscapes can inform the development of AVAS with functional meaning. Such studies could identify what acoustic dimensions (such as timbral shifts, rhythmic variations, or dynamic modulation) are most effective for conveying intent. By embedding these subtle yet informative cues into AVAS, systems could evolve from merely meeting regulatory requirements into genuinely communicative tools, enhancing not only detection but also pedestrians' ability to interpret and anticipate vehicle-driver behavior.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

C. Future Urban Soundscapes

A further consideration emerging from the EVA study relates to the future soundscape of urban environments as EVs proliferate. If all slow-moving traffic in a busy street comprises EVs with each emitting similar synthetic AVAS tones, pedestrians may face new challenges in interpreting and navigating this new acoustic environment. While individual AVAS designs can be assessed in isolation for detectability and compliance, the collective effect of multiple overlapping signals has not yet been meaningfully examined. AVAS tones that are designed to be individually detectable may, when layered together in dense traffic, create a homogenized and fatiguing auditory environment - ironically echoing the very concerns about noise pollution that traffic reduction campaigns have sought to address for decades in relation to ICE vehicles. Rather than enhancing safety, such soundscapes could overwhelm pedestrians with undifferentiated stimuli, reducing their ability to localize specific vehicles or interpret their behavior. The relative uniformity of AVAS (typically composed of tonal structures tied to speed) raises questions of differentiation. Will pedestrians be able to distinguish between vehicles approaching from different directions, or will multiple AVAS signals blur into a single background hum?

From the perspective of urban residents, these issues extend beyond safety into quality of life. Current regulatory frameworks are silent on the broader implications of large-scale AVAS deployment, focusing on minimum requirements for audibility without addressing the cumulative impact on the urban soundscape. If synthetic tones dominate city centers, there is a risk of creating a new form of noise pollution, one that is both uniform and omnipresent.

The challenge, therefore, is not only to ensure that AVAS provide sufficient detectability in isolation but also to consider their collective ecological validity in real-world environments. Solutions may involve introducing richer acoustic palettes that allow vehicles to remain distinct from one another or exploring adaptive sound designs that respond to environmental context (e.g., adjusting timbre or modulation in crowded versus quiet streets). Again, Müller, Forssén, and Kropp [65] provide empirical evidence of this problem occurring in current AVAS when two or three vehicles employed the same or similar tonal signal. Their research shows that participants frequently failed to localize individual AVAS sources, with tonal designs performing worst. Their findings reinforce concerns that uniform AVAS soundscapes in busy urban contexts may not only blur together but actively impair pedestrians' ability to navigate safely. Interestingly, their research demonstrates that ICE vehicles are more robust in these scenarios.

D. Bridging the Research–Standards Gap in AVAS

Despite the immediate safety implications of AVAS, there remains a striking gap between academic research and the development of standards, policy, and regulation. To date, AVAS standards have been primarily industry-driven, with efforts focused on measurable compliance parameters such as minimum sound levels and frequency components. While this has ensured regulatory clarity for manufacturers, it has left broader questions around perceptual effectiveness, functional meaning, and real-world pedestrian experience largely unaddressed. It is important to acknowledge that UNECE's QRTV process did invite and receive extensive submissions, which generated a substantial body of research data. However, much of this evidence came from third-party contributors, including a strong presence from industry and affiliated research groups. By contrast, the U.S. NHTSA directly commissioned and conducted structured human trials between 2011 and 2013, providing perhaps a more authoritative baseline for how pedestrians detect and interpret vehicle sounds.

UNECE's reliance on submitted data meant that the scope, comparators, and methodologies varied widely across trials, and it must be said that in some cases lacked transparency about critical parameters such as what ICE vehicles were used as benchmarks. Both processes ultimately converged on similar regulatory compromises, prioritizing minimum SPL and frequency thresholds over perceptual outcomes such as localization accuracy, notification distance, or anticipatory awareness. This institutional pathway is significant: it seems that authority-led trials yielded stronger and more consistent evidence, while UNECE's submissions-based model introduced variability that often diluted end-user perspectives. In this context, the likes of the EVA study may help to fill a gap in Europe by providing independent, systematically gathered end-user evidence that speaks directly to the perceptual challenges that AVAS regulations have struggled to resolve. Moreover, because ISO standards are a mechanism through which UNECE requirements are engaged and operationalized, the rigorous ISO and CEN standards processes may provide the most effective channel for incorporating academic research into the regulatory framework in future revisions. Yet, the opportunity to embed such evidence more systematically into the standards process remains limited, reflecting a broader structural disconnect between the kinds of long-term, psychometric studies undertaken in academia and the compliance-oriented priorities that drive regulatory development.

This imbalance points to a possible structural disconnect between academic research and standards development. Rigorous psychometric and perceptual studies are costly in terms of time and are methodologically demanding, yet are essential for evaluating AVAS effectiveness. Such work generally falls outside the remit of industry, where commercial pressures override such pursuits as evidenced in some of the QRTV minutes. This sort of research is more naturally suited to academia, particularly within publicly-funded programs capable of sustaining extended, exploratory investigation. However, while academic studies on EV detectability and sound design have been published, their insights have rarely been integrated into the iterative, real-world processes of standards development – mostly due to a

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

lack of direct academic involvement in the standards process itself.

The reasons for this disconnect are complex. Academic-reward structures often prioritize peer-reviewed publications, citations, and research grants, whereas active participation in standards-bodies is undervalued, despite its strict rigor and real-world impact. Moreover, the timelines and processes of international standards development do not always align neatly with academic cycles of funding and dissemination. The result is a sizeable, missed opportunity - for example, AVAS represents a case where academic evidence could and should be central but instead remains peripheral to the shaping of policy and regulation.

Addressing this gap requires not only greater academic engagement with standards organizations, but also a cultural shift in how research impact is measured. Contributions to standards - which undergo stringent peer review - demand technical precision and directly influence public safety. Therefore, standards contributions should be a recognized and legitimate form of academic output. In the case of AVAS, this alignment is particularly urgent. Without stronger academic participation, the field risks continuing along a path where compliance takes precedence over effectiveness, and where critical safety questions remain unresolved.

E. Limitations of the EVA Study

The EVA study, while providing valuable insights into pedestrian perceptions of AVAS, is subject to several limitations. First, the sample size, although sufficient for statistical analysis, was relatively modest. A larger and more diverse participant pool would strengthen the generalizability of findings, particularly across different demographic groups and geographic contexts. Second, while the survey design captured both quantitative Likert responses and qualitative open-ended insights, future studies could benefit from performance-based experimental designs that combine controlled acoustic testing with lived pedestrian experience. Such triangulation would provide richer evidence for linking perceptual data directly to measurable acoustic and psychometric features.

Another limitation is the scope of documentary analysis of regulatory processes. This paper reviewed selected UNECE QRTV meeting minutes [19] and linked them to subsequent Regulation No. 138, but it did not systematically examine the full corpus of UNECE QRTV and later WP.29 discussions, nor subsequent national transpositions of the regulation. A deeper meta-analysis of these records would allow for a clearer mapping of how specific perceptual concerns were introduced, debated, and either carried through or excluded from the regulatory text.

A further limitation lies in the absence of quantitative data on the degree of academic involvement in standards development. While this paper qualitatively notes that academic input was presented during UNECE's QRTV process but not always operationalized, future work should measure the extent of academic participation across different standards bodies. Such an analysis could also compare how various countries structure, incentivize, or fail to reward academic engagement in standards-making. This would help to clarify whether the observed gaps are structural features of the standards ecosystem itself or instead reflect national priorities and institutional cultures.

These limitations highlight the need for future work that scales up EVA-style studies, expands the review of regulatory archives, and systematically analyses the structural role of academia in standards development. Together, these steps would provide a more comprehensive evidence base for embedding pedestrian-centered perceptual research into future revisions of AVAS regulations.

IX. Conclusions

The EVA study highlights clear and consistent differences in how sighted and vision-impaired pedestrians experience the sounds of EVs. While current AVAS regulations ensure baseline audibility and compliance, they do not adequately address the functional needs of pedestrians who depend primarily on auditory cues. Across both groups, but especially among participants with vision-impairments, experiences of surprise and difficulty in anticipating vehicle actions point to fundamental gaps in the design and evaluation of AVAS.

These findings underscore three interlinked challenges for the future of AVAS: the absence of clear guidance on notification timing and effective distance; the tendency of many current implementations to prioritize branding over functional meaning; and the lack of consideration for the cumulative urban soundscape created by large-scale deployment of synthetic tones. Together, these gaps illustrate why compliance-based testing, while necessary, is insufficient for ensuring real-world safety and accessibility.

Evidence in the historical UNECE QRTV minutes shows that these gaps are not new concepts or phenomena. One particularly striking example came from D'Angelico (QRTV-06-08e, San Diego, 2011) [19], who focused on extrapolating how ICE vehicles naturally convey low-frequency modulations (40 Hz–130 Hz) that pedestrians intuitively associate with starting, stopping, or approaching. This, and other high-quality submissions - such as Pedersen's filter model incorporating perceptual hooks, and the European Association of Automotive Supplier's (CLEPA) focus on locatability - show that perceptual meaning was recognized as an essential dimension during the QRTV process, but was not ultimately operationalized in the regulation. Another example comes from the University of Duisburg-Essen (QRTV-09-02, Bonn, 2011) [19], which undertook a multi-phase study involving 240 participants, including blind and hearing-impaired groups. Their findings identified the 10–20 km/h "danger zone," where BEVs were consistently perceived as riskier than ICE vehicles despite only small physical SPL differences. Blind participants reported strong concerns about inaudibility, late detection, and potential social exclusion, with a clear preference for ICE-like cues over novel EV tones. Crucially, this study highlighted that perceived safety depends

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

as much on predictability and familiarity as on raw sound levels - a conclusion echoed in EVA's survey responses more than a decade later.

It is also worth recognizing that such early contributions in 2010 and 2011 flagged issues such as quiet ICE comparators, e-bike inclusion, and notification distance, yet many of these concerns were only partially reflected in the published Regulation No. 138, which as already mentioned, centered primarily on SPL and frequency thresholds. The EVA results therefore confirm unresolved issues that were identified in UNECE's formative years but not fully carried through into regulation, highlighting the continuing need for pedestrian-centered evidence to guide future revisions.

Equally important is the structural gap between academic research and the standards-making processes that eventually help shape AVAS regulation. To date, regulation has been largely industry-driven, with limited incorporation of empirical insights from up-to-date perceptual and psychometric research. Yet it is precisely this kind of research that is needed to ensure that AVAS functions as an effective communication tool rather than a mere regulatory requirement. Strengthening academic participation in standards would not only bring additional scientific rigor to policy design but would also ensure that all pedestrian groups are properly represented in regulatory and policy frameworks.

The EVA study, in a small way, illustrates the value of bridging these worlds. It demonstrates how collaboration between public bodies, research funders, and academia can connect data collection to lived pedestrian experiences while framing those insights within the context of standards, regulation, and policy. Such partnerships show how publicly funded research can address immediate safety-critical issues and help close the gap between research evidence, policy, and applied implementation.

X. Future Work

Future work on EVA will move in several complementary directions. The first is an expanded meta-analysis of regulatory history, examining the full record of QRTV and WP.29 submissions and minutes, and mapping how these inputs were distilled in the final text of UNECE Regulation No. 138 and subsequent EU legislation. Particular attention will be given to how these influenced the ISO and CEN standards, and how they themselves have fed back into regulatory updates. This historical–regulatory mapping will clarify how early perceptual concerns were translated, diluted, or excluded in the journey from submission to regulation or standard.

Equally, there is a need to develop and validate perceptual metrics that can be operationalized within standards compliance testing. One promising pathway is to revisit earlier proposals raised during the QRTV sessions, such as locatability hooks and distance-based perceptual measures, and explore how these can be translated into measurable standards today. By reaching out to the researchers and organizations that contributed such ideas in those formative years, academia can help refine and extend these approaches into robust, standardized protocols with a decade of additional data now available. In this way, future revisions of AVAS regulation could move beyond compliance with minimal thresholds and begin to reflect more of a pedestrian-centered framework grounded in "perceptual effectiveness".

In addition to performance-based user studies, it will be important for the next phase of EVA to expand its methodological scope. Beyond surveys, it will be important to also conduct interviews and focus-group sessions with pedestrians across various demographics to allow for richer, qualitative engagement with user-perspectives. Braun and Clarke's guidelines [66] will be applied in this context to generate themes grounded in participants' lived-experiences, providing deeper insights into how AVAS is interpreted in real urban environments.

A further direction in EVA will focus on academic engagement in standards development, since many of the problems highlighted in this article may perhaps be a symptom (at least in part) of a disconnect between academic outputs and standards development. A systematic analysis of the extent, type, and impact of academic contributions to AVAS standards will be undertaken, contrasting them with more industry-led inputs. This will include exploring how psychoacoustic and perceptual data can be translated into measurable, compliance-ready metrics, bridging the gap between complex psychometric research and regulatory enforcement. Particular emphasis will be placed on whether universal-design principles, as promoted by public authorities and disability organizations, can provide a framework for embedding the perspectives of end-users and stakeholders more consistently into standards processes.

Looking further ahead, future AVAS design may be enhanced through the integration of artificial intelligence and context-aware technologies. An AI-enabled AVAS could dynamically adapt its signals based on environmental noise, pedestrian density, time of day, or the presence of vulnerable users, ensuring notifications are both perceptually effective and socially considerate. Such systems could form part of broader connected-vehicle and smart city infrastructures, combining acoustic cues with multimodal safety-channels. Rather than static compliance artefacts, AVAS could evolve into adaptive, intelligent communication systems that anticipate and respond to real-world contexts, providing a more inclusive and effective safeguard for pedestrians as urban mobility transforms.

Acknowledgment

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 23/PSF/12154.

Taighde Éireann
Research Ireland

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

REFERENCES

[1] F. Neff, D. Fitzpatrick, G. Craddock, C. Cheevers, and A. Conway, "Assessing the Effectiveness of Acoustic Vehicle Alerting Systems (AVAS) for Pedestrians with Visual Disabilities," in *the Tenth International Conference on Universal Accessibility in the Internet of Things and Smart Environments*, IARIA Congress, Nice, France, 2025.

[2] P. J. Edwards, S. E. Moore, and C. Higgins, "Pedestrian safety on the road to net zero: cross-sectional study of collisions with electric and hybrid-electric cars in Great Britain," *J Epidemiol Community Health,* vol. 78, no. 8, pp. 487-492, 10 July 2024.

[3] Y. Cai, "Pedestrian auditory perception of approaching vehicles from behind in shared space: The impact of quietness of electric vehicles," *Transportation research procedia,* vol. 78, pp. 594-601, 2024.

[4] F. Mocholí Belenguer, A. Martínez-Millana, F. S. Castells Ramón, and A. Mocholí-Salcedo, "The Effectiveness of Alert Sounds for Electric Vehicles Based on Pedestrians' Perception," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 4, pp. 2956-2965, April 2022.

[5] J. Krautwurm, D. Oberfeld-Twistel, T. Huisman, M. Mareen Maravich, and E. Altinsoy, "Quality Comparison of Dynamic Auditory Virtual-Reality Simulation Approaches of Approaching Vehicles Regarding Perceptual Behavior and Psychoacoustic Values," *Acoustics,* vol. 7, no. 1, 2025.

[6] S. S. Dahl, J. Oddershede, J. Brunskog, C. P. Volk, and M. C. Van Eeckhoutte, "Noise differences between electric and conventional cars at low speeds," in *Proc. DAS|DAGA 2025*, pp. 1483–1486, 2025, doi: 10.71568/dasdaga2025.595.

[7] M. Roan, L. Neurauter, D. Moore, and D. Glaser, "Assessing the Ability of Pedestrians to Safely Hear Approaching Electric Vehicles," *Sound and Vibration,* vol. 51, no. 10, pp. 6-13, 2017.

[8] M. Pardo-Ferreira, J. Torrecilla-García, C. Heras-Rosas, and J. Rubio-Romero, "New Risk Situations Related to Low Noise from Electric Vehicles: Perception of Workers as Pedestrians and Other Vehicle Drivers," *International Journal of Environ. Res. Public Health,* vol. 17, no. 6701, 2020.

[9] D. Oberfeld and T. Huisman, "Pedestrians' street-crossing decisions compared between conventional and electric vehicles," in *NOISE-CON proceedings*, New Orleans, USA, 2024.

[10] M. Wessels, S. Kröling, and D. Oberfeld, "Audiovisual time-to-collision estimation for accelerating vehicles: The acoustic signature of electric vehicles impairs pedestrians' judgments," *Transportation Research Part F - Traffic Psychology and Behaviour,* vol. 91, pp. 191-212, 2022.

[11] L. Müller and W. Kropp, "Auralization of electric vehicles for the perceptual evaluation of acoustic vehicle alerting systems," *Acta Acustica,* vol. 8, no. 27, 2024.

[12] P. R. DeLucia, D. Oberfeld, J. K. Kearney, M. Cloutier, A. M. Jilla, A. Zhou, S. Trejo Corona, J. Cormier, and A. Taylor, "Time-to-Collision Estimation With Age-Related Macular Degeneration Using Visual and Auditory Cues: Which Cues are Most Important?," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 2025.

[13] M. Roan, M. Neurauter, D. Moore, and D. Glaser, "Electric Vehicle Detectability: a Methods-Based Approach to Assess Artificial Noise Impact on the Ability of Pedestrians to Safely Detect Approaching Electric Vehicles," *SAE International Journal of Vehicle Dynamics, Stability, and NVH,* vol. 1, no. 2, pp. 352-361, 2017.

[14] T. Berge, "Experience and Perception of AVAS on Electric Vehicles in Norway," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings, InterNoise 18*, Chicago, USA, 2018.

[15] Eurostat, "Passenger cars in the EU - Statistics Explained," European Commission, Brussels, 2024.

[16] European Commission, "(EU) 2017/1576 amendment of Regulation (EU) No 540/2014 of the European Parliament and of the Council as regards the Acoustic Vehicle Alerting System requirements for vehicle EU-type approval," *Official Journal of the European Union,* vol. L 239, no. Sept., pp. 3–6, 2017.

[17] European Parliament and Council, "Regulation (EU) No. 540/2014 of 16 April 2014 on the sound level of motor vehicles and of replacement silencing systems," Official Journal of the European Union, Brussels, 2014.

[18] United Nations Economic Commission for Europe (UNECE), "Regulation No. 138: Uniform provisions concerning the approval of quiet road transport vehicles with regard to their reduced audibility," UNECE, Geneva, Switzerland, 2016.

[19] UNECE, "QRTV Previous Sessions - UNECE Transport Division Wiki," [Online]. Available: https://wiki.unece.org/display/trans/QRTV+Previous+Sessions. [Accessed 10 09 2025].

[20] U. T. Division, "World Forum for Harmonization of Vehicle Regulations (WP.29)," [Online]. Available: https://unece.org/transport/vehicle-regulations/world-forum-harmonization-vehicle-regulations-wp29. [Accessed 18 August 2025].

[21] U.S. Department of Transportation, National Highway Traffic Safety Administration, "Federal Motor Vehicle Safety Standards; Minimum Sound Requirements for Hybrid and Electric Vehicles. Proposed Rule," *49 CFR Parts 571 and 585, Docket No. NHTSA-2011-0100, Fed. Reg.,* vol. 78, no. 9, pp. 2797–2817, 2013.

[22] A. Fiebig, "Electric vehicles get alert signals to be heard by pedestrians: Benefits and drawbacks," *Acoustics Today,* vol. 16, no. 4, pp. 20-28, 2020.

[23] National Highway Traffic Safety Administration, *Federal Motor Vehicle Safety Standard No. 141: Minimum sound requirements for hybrid and electric vehicles Code of Federal Regulations, Title 49, Part 571.141,* Washington, D.C., USA: National Highway Traffic Safety Administration, 2018.

[24] Standardization Administration of China (SAC) and the State Administration for Market Regulation (SAMR), *GB/T 37153-2018 - Acoustic Vehicle Alerting System of Electric Vehicles Running at Low Speed,* Beijing: Standardization Administration of the People's Republic of China, 2018.

[25] International Organization for Standardization, *ISO 16254:2024 - Acoustics - Measurement of sound emitted by road vehicles of category M and N at standstill and low speed operation - Engineering method,* Geneva, Switzerland: ISO, 2024.

[26] A. Hastings, C. Scarpone, B. Samiljan, L. Garay-Vega, and J. K. Pollard, "Acoustic Characteristics of Hybrid Electric Vehicles and Safety of Blind Pedestrians," *Transportation Research Record,,* vol. 2248, no. 08, pp. 61-67, 2011.

[27] U. Sandberg, L. Goubert, and P. Mioduszewski, "Are vehicles driven in electric mode so quiet that they need acoustic warning signals," in *Proceedings of 20th International Congress on Acoustics, ICA 2010*, Sydney, Australia, 2010.

[28] A. Fortino, L. Eckstein, J. Viehöfer, and J. Pampel, "Acoustic Vehicle Alerting Systems (AVAS) - Regulations, Realization and Sound Design Challenges,," *SAE International Journal of Passenger Cars—Mechanical Systems,* vol. 9, no. 3, pp. 995-1003, 2016.

[29] A. Pyzik and P. Larsson, "Acoustic vehicle alerting system and method". United States of America Patent 11,904,772 B2, 20 Feb 2024.

[30] E. M. Fontecha, G. Waters, and T. R. Stuart, "Performance and acceptability of the TfL Urban Bus Sound (AVAS)," in *INTER-*

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

NOISE and NOISE-CON Congress and Conference, InterNoise22, Glasgow, Scotland, 2022.

[31] B. Kullukcu, M. Engel, and S. Atamer, "Psychoacoustic Modelling of AVAS Sounds: Consumer-Centric Semantic Attribute Development for Electric Vehicles," in *11th Convention of the European Acoustics Association, Forum Acusticum, EURONOISE*, Málaga, Spain, 2025.

[32] T. Walton, A. J. Torija, and A. S. Elliott, "Development of electric scooter alerting sounds using psychoacoustical metrics," *Applied Acoustics,* vol. 201: 109136, December 2022.

[33] M. A. Pallas, M. Capatina, A. Le Bellec, S. Bianchetti and J. Cesbron, "Environmental noise from alerting signals (AVAS) of electric vehicles: comparison with noise prediction models at urban speeds,," in *10th Convention of the European Acoustics Association Forum Acusticum 2023*, Turin, Italy, 2023.

[34] F. Gutierrez, X. Robin, and A. Charon, "Striking the right note: balancing AVAS in tomorrow's soundscapes," in *NOISE-CON*, New Orleans, USA, 2024.

[35] M. Hsieh, H. Chen, M. Tong, and C. Yan, "Effect of Environmental Noise, Distance and Warning Sound on Pedestrians' Auditory Detectability of Electric Vehicles," *International Journal of Environmental Research and Public Health (IJERPH),* vol. 18, no. 17: 9290, 2 September 2021.

[36] D. Oberfeld, M. Wessels, and T. Huisman, "Auditory and audiovisual time-to-collision estimation and road-crossing decisions," in *10th Convention of the European Acoustics Association*, Turin, Italy, 2023.

[37] R. Sanz-Segura and E. Manchado-Pérez, "Does electric-mobility sound futuristic? conceptualization design strategies for the auditory vehicle alerting system (avas)," *DYNA,* vol. 96, no. 1, pp. 26-29, 2020.

[38] S. Lenzi, J. Sádaba, and P. Ciuccarelli, "A designerly approach to the sonification of electric vehicles," in *Design, Research, Society: DRS2022*, Bilbao, Spain, 2022.

[39] S. Lenzi and G. D'Amico, "Sound for Electric Vehicles," in *Audio Branding Yearbook*, Berlin, Germany, Nomos Verlag, 2010.

[40] A. Zeitler, "Psychoacoustic requirements for warning sounds of quiet cars," *SAE International Journal of Passenger Cars - Electronic and Electrical Systems,* vol. 5, no. 2, pp. 572–578, 2012.

[41] S. K. Lee, S. M. Lee, T. Shin, and M. Han, "Objective evaluation of the sound quality of the warning sound of electric vehicles with a consideration of the masking effect: Annoyance and detectability,," *International Journal of Automotive Technology,* vol. 18, no. 4, pp. 699–705, 2017.

[42] K. Yamauchi, T. Sano, S. Hasegawa, F. Tamura, and Y. Takeda, "Detectability and hearing impression of additional warning sounds for electric or hybrid vehicles," in *INTERNOISE 2014 - 43rd International Congress on Noise Control Engineering: Improving the World Through Noise Control*, Melbourne, Australia, 2014.

[43] N. Misdariis, A. Cera, R. Barthelemy, L. Worms, and L.-F. Pardo, "Collaborative sound design for automotive applications - the sound for electric vehicle as an emblematic use case," in *NOISE-CON*, Nantes, France, 2024.

[44] N. Misdariis and A. Cera, "Knowledge in Sound Design: The Silent Electric Vehicle—A Relevant Case Study," in *Proceedings of the Conference on Design and Semantics of Form and Movement - Sense and Sensitivity*, Eindhoven, Netherlands, 2017.

[45] T. Belschner and M. Bodden, "Comprehensive automotive active sound design-part 2: Operational sounds and brand sound.," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Melbourne, Australia, 2014.

[46] M. Carron, F. Dubois, N. Misdariis, C. Talotte, and P. Susini, "Designing sound identity: providing new communication tools for building brands "corporate sound"," in *9th Audio Mostly: A Conference on Interaction With Sound*, New York, USA, 2014.

[47] L. Zattra, N. Misdariis, F. Pecquet, N. Donin, and D. Fierro, "Analysis of Sound Design Practices [ASDP]: Research Methodology," in *XXII Colloquio di Informatica Musicale (Congres d'Informatique Musicale CIM)*, Udine, Italy, 2018.

[48] M. Liljedahl and J. Fagerlönn, "Methods for sound design: a review and implications for research and practice," in *5th Audio Mostly Conference: A Conference on Interaction with Sound*, New York, USA, 2010.

[49] N. Misdariis, A. Gruson, and P. Susini, "Detectability study of warning signals in urban background noises: A first step for designing the sound of electric vehicles," *Journal of the Acoustical Society of America,* vol. 19, no. 1, 2013.

[50] R. Stock, "Sonic e-mobility: traffic noise, sound-producing electric vehicles, and blind pedestrians," *Mobilities,* vol. 20, no. 3, pp. 410–426, 05 Jan 2025.

[51] M. Allman-Ward, R. O. Williams, T. Heinz, and M. Demontis, "Designing and delivering the right sound for quiet vehicles,," in *Inter-noise 2014*, Melbourne, Australia, 2014.

[52] C. Kirsch, A. Baki Kocaballi, A. Johnston, A. Naweed, and I. Stevenson, "Co-Designing the Sound of Safety: Embracing Complexity in the Acoustic Vehicle Alerting System Sound for Zero Emission Buses," *Human Factors and Ergonomics in Manufacturing & Service Industries,* vol. 35, no. 3, 2025.

[53] P. Fink, S. Doore, X. Lin, M. Maring, P. Zhao, A. Nygaard, G. Beals, R. Corey, R. Perry, K. Freund, V. Dimitrov, and N. Giudice, "The Autonomous Vehicle Assistant (AVA): Emerging technology design supporting blind and visually impaired travelers in autonomous transportation," *International Journal of Human-Computer Studies,* vol. 179, 2023.

[54] F. M. Belenguer, A. Martinez-Millana, F. S. C. Ramón, and A. Mocholi-Salcedo, "The effectiveness of alert sounds for electric vehicles based on pedestrians' perception," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 4, pp. 2956-2965, 2020.

[55] J. F. Petiot, B. G. Kristensen, and A. Maier, "How should an electric vehicle sound? User and expert perception," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Portland, Oregon, USA, 2013.

[56] E. Zwicker and H. Fastl, Psychoacoustics: Facts and Models, Berlin, Germany: Springer, 1999.

[57] N. Yasui and M. Miura, "Design of alert sound for electric vehicle based on fluctuation strength for amplitude fluctuation," in *INTER-NOISE and NOISE-CON Congress and Conference, InterNoise23*, Chiba, Japan, 2023.

[58] K. Yamauchi, "Effectiveness of additional warning sounds for quiet vehicles in urban noise environment," *Journal of the Acoustical Society of America,* vol. 140, no. 4, November 2016.

[59] T. Souaille, "An interactive bi-objective optimisation process to guide the design of electric vehicle warning sounds," *Design science,* vol. 8, Jan 2022.

[60] M. Calloni, R. Hadjit, P. Salvekar, and C. Musser, "Harmonizing Safety Regulations and Perception: A Simulation Methodology for AVAS System Design," *SAE technical paper series,* Vols. No. 2024-01-2915, June 2024.

[61] International Organization for Standardization, *ISO 10844:2021 – Acoustics — Specification of test tracks for measuring noise emitted by road vehicles,* Geneva, Switzerland: ISO, 2021.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[62] National Disability Authority, "Advice Paper on Disability Language and Terminology," 2023. [Online]. Available: https://nda.ie/uploads/publications/NDA-Advice- Paper-on-Disability-Language-and-Terminology.docx. [Accessed April 2025].

[63] H. McHugh and F. Dolan, "Guidelines for Good Adult Literacy Work," National Adult Literacy Agency, Dublin, Ireland, 2012.

[64] F. Neff, D. Fitzpatrick, G. Craddock, and C. Cheevers, "Making Electric Vehicles Audible: Perception and Inclusivity in Acoustic Vehicle Alerting System Design," in *30th Association for the Advancement of Assistive Technology in Europe (AAATE)*, Nicosia, Cyprus, 2025.

[65] L. Müller, J. Forssén, and W. Kropp, "Auditory localization of multiple stationary electric vehicles," *The Journal of the Acoustical Society of America,* vol. 157, no. 3, pp. 2029–2041, 2025.

[66] V. Braun and V. Clarke, "Using thematic analysis in psychology.," *Qualitative Research in Psychology,* vol. 3, no. 2, pp. 77–101, 2006.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Leveraging Content and Structural Dynamics of Discourse for Rumor Verification

Gibson Nkhata, Susan Gauch
Department of Electrical Engineering and Computer Science
University of Arkansas
Fayetteville, AR 72701, USA
e-mails: {gnkhata, sgauch}@uark.edu

***Abstract*—The rapid spread of misinformation on social media platforms has heightened the need for effective rumor verification models. Traditional approaches to rumor verification primarily rely on textual content and transformer-based embeddings, but they often fail to incorporate conversational dynamics and stance evolution, limiting their effectiveness, necessitating the need for robust and interpretable rumor verification systems. This study introduces a novel framework that integrates semantic content, structural dynamics, and stance distribution features into a unified model for predicting rumor veracity. By employing hierarchical sequencing strategies, including Breadth-First Search, Depth-First Search, and temporal traversal, for structural dynamics, this work effectively captures both global and local relationships within conversational threads. This work further leverages an attention-based encoder to aggregate content embeddings, stance distributions, and reply-level features, overcoming limitations of sequence truncation and underutilized conversational structures found in existing methods. Experimental results on benchmark datasets, including SemEval-2017, RumorEval-2019, and PHEME, demonstrate that our approach achieves state-of-the-art performance, significantly improving Macro-F1 scores and accuracy over competing models. Ablation studies confirm the critical contributions of hierarchical encoding, stance aggregation, and attention mechanisms to the model's success. Thus, this work sets a new standard for efficient, interpretable, and scalable rumor verification, offering promising directions for mitigating misinformation on social media platforms.**

***Index Terms*—Rumor verification; stance-conditioned modeling; social media misinformation; early detection; embedding aggregation.**

I. Introduction

This paper extends our previous work, stance-conditioned modeling for rumor verification [1]. Specifically, the prior work focused on integrating source and reply post embeddings from Bidirectional Encoder Representations from Transformers (BERT) [2] with stance labels, encoded through Bidirectional Long Short Term Memory (BiLSTM) [3], for rumor classification.

Social media platforms play an essential role in information dissemination, news sharing, and communication in the modern Internet era [4][5], yet they also foster the proliferation of misinformation, which poses a significant challenge to societal trust, with far-reaching implications for public health, politics, and safety [6][7]. Therefore, automated rumor verification systems have emerged as crucial tools to combat misinformation, relying on machine learning and natural language processing (NLP) techniques [8][9]. Despite progress, existing systems face key limitations in effectively utilizing the entirety of conversational threads, particularly in capturing both semantic and structural dynamics.

For example, [8] segment lengthy threads into shorter subthreads and use BERT to individually encode each subthread. A global model layer is then applied to integrate the representations of all subthreads, with constraints on the maximum number of posts per thread and the maximum number of subthreads per thread. While existing work predominantly utilize pre-trained Large Language Models (LLMs) like BERT, these methods face inherent limitations. One major challenge is the sequence length constraint of LLMs, often used to encode full conversational threads. LLMs truncate inputs or subdivide threads into smaller chunks, resulting in the omission of replies that contain critical semantic information. Moreover, structural information, such as the hierarchy of replies, stance distributions (e.g., Support, Deny, Query, Comment), and the temporal sequencing of posts' features, is either underutilized or entirely neglected in a discourse. Such limitations hinder the development of robust models that can comprehensively analyze discourse for accurate rumor veracity prediction.

Recent advances in rumor verification, mostly breaking event rumor detection, has shown that stance-separated models have improved veracity prediction by analyzing distinct stance categories during breaking news events [10]. Furthermore, multi-task frameworks integrating stance classification with rumor detection have shown improved performance by leveraging shared features across tasks [11]. Additionally, modern graph-based enhancements include stance distributions as node-level features in graph attention mechanisms, underscoring its importance in contextualizing relationships within threads [12].

We propose a novel approach that integrates content embeddings, stance aggregation, and structural information into a unified model for rumor verification. By integrating hierarchical and temporal sequencing of discourse into a unified framework, this research bridges the gap between content-based and structure-aware rumor verification. We use stance-based posts' aggregation and graph-based traversal techniques for post-level vector encoding to ensure a holistic representation of conversational threads, overcoming the limitations of sequence truncation and incomplete conversational thread

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

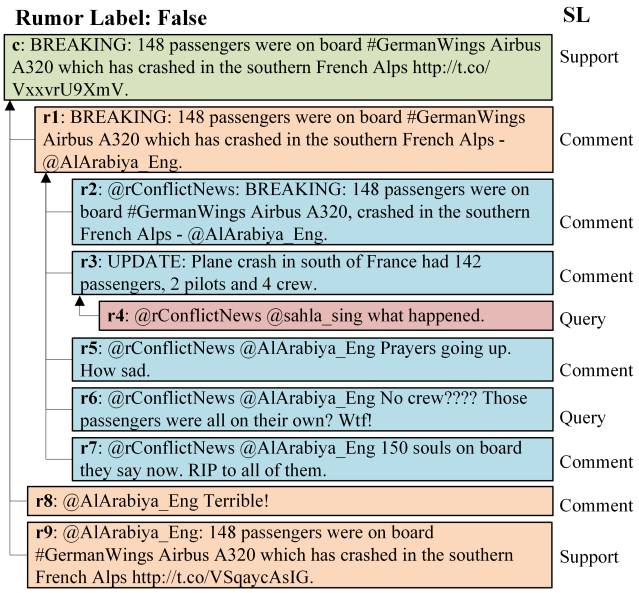


Fig. 1. A sample thread C with a false veracity label. SL stands for Stance Labels.

utilization. While methodologies such as GACN [13] and SAMGAT [12] focus on graph-based models, our approach departs from explicit graph neural networks and instead employs graph-inspired traversal strategies for sequencing. This provides a lightweight yet effective way to capture structural and temporal dynamics without the computational overhead of full graph-based models.

Furthermore, this study explores the task of early rumor detection, focusing on identifying and assessing the veracity of emerging rumors in real-time as they propagate online. By detecting rumors at an early stage, this approach aims to mitigate the rapid spread of misinformation, enabling timely interventions and fact-checking before false narratives gain widespread traction. Figure 1 presents a sample discourse, showcasing how stances evolve. Our work sets a new standard for efficient and interpretable systems that leverage discourse dynamics to combat misinformation.

The key contributions of this work are:

- **Rich Feature Representation**: Our framework integrates embeddings of a source post and replies grouped by stance type. Hierarchical reply levels and stance distributions are explicitly modeled as input features, enabling the model to learn from structural and temporal dynamics.
- **Hierarchical Sequencing of Discourse Trees**: This work models each conversation thread as a discourse tree and sequences posts using Breadth-First Search (BFS), Depth-First Search (DFS) traversals, and temporal sequencing, systematically organizing reply levels and robustly representing the hierarchical relationships within the discourse.
- **Enhanced Model Architecture**: We present an attention-based encoder that processes posts' embeddings, stance distributions, and reply-level vectors derived from graph traversal strategies.
- **Rigorous Experimental Validation**: The approach is evaluated on competitive and publicly available datasets, demonstrating significant improvements over state-of-the-art (SOTA) models in metrics such as Macro-F1 score and accuracy. Furthermore, we evaluate the model's ability to detect rumors at an early stage of the conversation, highlighting its real-world applicability for misinformation mitigation

The remainder of this paper unfolds as follows. Section II reviews the existing literature on rumor verification, and Section III delves into a comprehensive description of our approach. Section IV demonstrates experiments and provides a discussion of results. Finally, Section V concludes the paper.

II. Related Work

A rumor is defined as a widely circulated piece of information whose veracity is uncertain [14][15]. Rumours appear credible but lack immediate verification and often provokes skepticism, thus prompting users or automated systems to seek confirmation of its truthfulness. The field of rumor verification has witnessed significant advancements in recent years, driven by the need to combat the spread of misinformation on social media platforms. Most existing studies can be categorized into content-based approaches, structure-aware models, and methods integrating temporal and hierarchical features.

A. Content-Based Approaches

The early methods of rumor verification focused primarily on textual content using traditional machine learning and NLP techniques [16, 17, 18]. With the rise of deep learning, numerous models have been applied to the task of rumor verification. Early deep learning studies predominantly utilized Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), to extract information from rumor conversations. These approaches typically arrange tweets chronologically to represent the entire conversation's features [19, 20, 21]. However, such methods face challenges in effectively capturing long-range dependencies within sequences and rely on sequential processing. Therefore, transformer-based pre-trained models like BERT and its variants became the backbone of many rumor verification systems [9]. Nevertheless, these models suffer from sequence length limitations, often truncating crucial replies or ignoring their hierarchical relationships within threads.

B. Structure-Aware Models

Several studies have incorporated discourse structures into their models, recognizing the importance of structural features [22, 23, 24, 25]. For instance, [22] employed tree-structured recursive neural networks to model hierarchical reply relationships in Twitter (rebranded as X) threads. More recently, graph-based approaches like SAMGAT [12] have introduced multilevel graph attention networks to integrate the semantic relationships between posts and their replies. In addition, a sequence graph network [26] has modeled conversations

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

as graphs and utilized a graph attention mechanism to process the interactions within them. While graph neural networks provide a powerful tool for capturing structural and semantic interactions, they often involve significant computational overhead and lack explicit focus on hierarchical sequencing.

C. Temporal and Hierarchical Features

Temporal dynamics and reply hierarchies have also been explored to enhance rumor verification systems. [27] highlighted the importance of temporal sequencing in understanding the progression of online discussions. Similarly, methods like DynamicGCN [28] use temporal information to model the propagation of rumors. Still, these approaches typically focus on global temporal patterns, overlooking the local reply-level interactions and structural nuances

To this end, existing rumor verification work often adopts flat sequence-based approaches or computationally intensive graph-based models, leaving key gaps in their ability to capture discourse dynamics holistically. Flat-sequence approaches fail to represent hierarchical and temporal relationships, while graph-based models focus on learning node and edge interactions but come with significant computational overhead. We further notice that many existing methods overlook the integration of nuanced structural features, such as reply hierarchies and stance distributions, alongside temporal patterns of replies. This research bridges these gaps by employing lightweight graph traversal strategies and temporal sequencing to sequence hierarchical relationships in discourse trees. Using these strategies, we create a solution that effectively encodes hierarchical, temporal, and semantic features without relying on computationally heavy graph-based neural networks, providing a scalable, interpretable, and comprehensive technique for capturing the interplay between content and structure within discourse.

III. Methodology

Our model consists of four main components: 1) **Post embedding representation**: BERT extracts contextual embeddings for the source and reply posts. 2) **Structural features extraction**: We encode hierarchical levels and stance distribution. 3) **An attention-based encoder**: The post embeddings and structural features are integrated. 4) **A decoder**: used for rumor prediction. Figure 2 illustrates our methodology. We begin by formally defining the problem of rumor verification, followed by a detailed explanation of the methodology.

A. Problem Definition

The dataset for rumor detection is defined as $E = \{e_1, e_2, \ldots, e_n\}$, where each e_i corresponds to a unique rumor event and n refers to the total number of rumor events. Each event e_i consists of (c, y, R, s), where c represents the claim, and y denotes the veracity label associated with c, such that $y \in \{\text{true, false, unverified}\}$ rumor. The responses in the discourse for claim c are arranged chronologically as $R(c) = \{r_1, r_2, \ldots, r_m\}$, where m indicates the number of posts r responding to a particular claim c. s encapsulates the structural information of the discourse tree. To this end, this task is formulated as a supervised classification problem, where the goal is to learn a function $f : E \to y$ that predicts the veracity label of each event e_i.

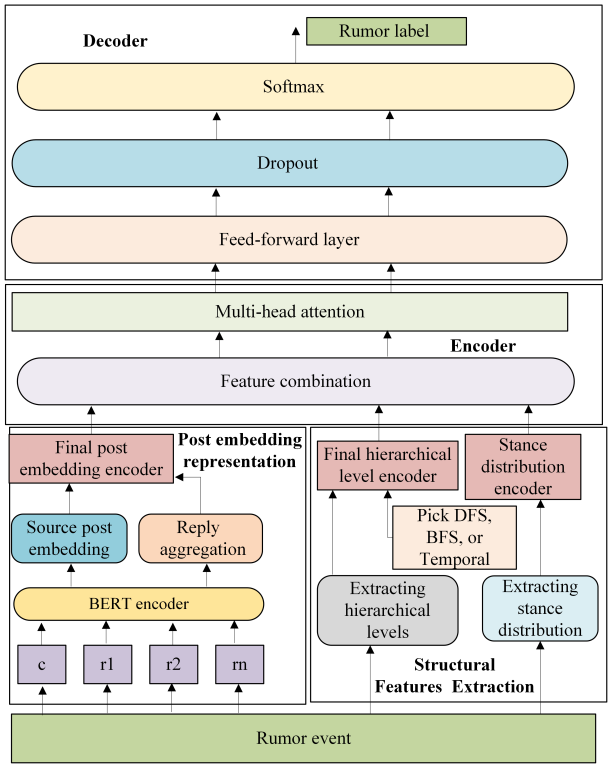


Fig. 2. The model framework.

B. Post embedding representation

Each post x_i (both c and r_i) in a conversational thread is firstly tokenized and passed through a pre-trained language model (e.g., BERT). This produces contextual embeddings for each token in the post. A pooling strategy is then applied to aggregate these token embeddings into a single vector representing the entire post as

$$\mathbf{e}_i = \text{BERT}(x_i) = \frac{1}{T}\sum_{t=1}^{T} h_t, \tag{1}$$

where h_t represents the hidden state at position t of a given post, and T is BERT input sequence length.

The stance of each reply plays a critical role in the discourse's overall meaning. Following prior work on stance aggregation [7][17][29], we categorize all replies in a thread stance labels (Support, Deny, Query, Comment). For each stance type, embeddings of all replies sharing that stance are aggregated to create a single vector representing the semantic contribution of that stance to the discourse. Let $V_{\text{stance}} \subseteq V$ be the set of nodes (replies) in the thread that share a particular stance. The aggregated embedding for this stance is:

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

$$\mathbf{e}_{\text{stance}} = \frac{1}{|V_{\text{stance}}|} \sum_{v_i \in V_{\text{stance}}} \mathbf{e}(v_i), \tag{2}$$

If no replies belong to a particular stance, a zero vector is used, following related practices in hierarchical discourse modeling [17].

After aggregating embeddings for all four stances, these vectors are concatenated with the embedding of the source post to create a composite feature vector for a rumor event:

$$f = [e_{src}, e_{support}, e_{deny}, e_{query}, e_{comment}] \tag{3}$$

In rumor verification tasks, aggregating embeddings efficiently and meaningfully is crucial to capturing both semantic content and the structural dynamics of conversations. Moreover, as illustrated in Figure 1, posts sharing the same stance tend to convey similar semantic information. Thus, we propose that aggregating these posts can provide a comprehensive representation of the semantic information within a conversational thread. Additionally, stance aggregation ensures dimensionality reduction, where aggregating embeddings by stance reduces computational overhead while preserving essential information [6], thereby overcoming the limitations of thread sequence truncation. Finally, aggregation enforces scalability, ensuring that the model handles long threads without exceeding memory constraints [17][30].

f from (3) is then concatenated with hierarchical and structural information to form the input features for the model.

C. Structural Features Extraction

Building on recent advances in discourse modeling [12][28], structural and temporal features are incorporated to enhance a composite representation. First, to capture the hierarchical relationships within conversational threads, we represent the discourse as a tree structure, where each node corresponds to a post, and edges represent reply relationships. Second, for traversal and sequencing, traversal techniques—BFS, DFS, and temporal sequencing—are applied to sequence replies and encode their hierarchical levels into a structured feature vector. Finally, we integrate stance distribution information.

Let the discourse tree $G = (V, E)$ represent a conversational thread. V is the set of nodes, where $v_i \in V$ corresponds to a post. E is the set of directed edges, where $e_{ij} \in E$ indicates that v_i is a reply to v_j. The root node v_r represents the source post c; the other nodes v_i are replies r_i.

1) Encoding hierarchical levels: The hierarchical level $\ell(vi)$ of each node v_i is defined recursively, $\ell(vi) = \ell(v_p)+1$, where v_p is the parent of v_i. Each node v_i is assigned a reply-level vector based on its position in the tree. The root node is by default promoted to level 1. For a tree with maximum depth D, the vector $h(v_i)$ is:

$$h(v_i) = [0, \ldots, 0, 1, 0, \ldots, 0], \tag{4}$$

where the $\ell(vi)$-th position is set to 1, and all others are 0; $h(v_i)$'s dimension is D. Therefore, we sequence $h(v_i)$ vectors by separately adopting the following three tree traversal strategies.

1) **Breadth-First Search:** BFS sequences nodes level by level, starting from the root node v_r. At each level, all nodes are processed before moving to the next level. This traversal preserves the hierarchy in a breadth-wise manner when applying the hierarchical level $\ell(v_i)$. For example, considering the discourse tree in Figure 1, the breadth-first order of the propagation structure would be $[c, r_1, r_8, r_9, r_2, r_3, r_5, r_6, r_7, r_4]$.
2) **Depth-First Search:** DFS explores as far as possible along each branch before backtracking. This traversal preserves the depth-first order of replies and captures long chains of interactions. This sequence highlights the reply chains within a thread. Referring again to Figure 1, the depth-first order of the propagation structure would be $[c, r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8, r_9]$.
3) **Temporal sequencing:** Reply timestamps are appended to capture the evolution of discussions over time [28]. To integrate the time dimension, each node v_i is assigned a timestamp $t(v_i)$, representing the post's creation time. The replies are then sorted by $t(vi)$ within each hierarchical level. Given two nodes (v_i, v_j),

$$\text{Order}(v_i, v_j) = \begin{cases} -1 & \text{if } t(v_i) < t(v_j), \\ 1 & \text{if } t(v_i) > t(v_j). \end{cases} \tag{5}$$

This ordering ensures that replies are chronologically processed while maintaining their hierarchical structure.

Whether BFS, DFS, or temporal sequencing is used, it results in a hierarchical vector $\mathbf{h}_{\text{input}}$.

2) Stance Distribution: We subsequently integrate a normalized vector representing the proportion of replies associated with each stance category. This design is inspired by prior work, such as SAMGAT [31], which employs graph-based aggregation to capture reply distribution patterns. The resulting feature serves as a global descriptor of the discourse's stance dynamics and is formally expressed as a normalized vector:

$$\mathbf{s} = \left[\frac{|V_{\text{support}}|}{|V|}, \frac{|V_{\text{deny}}|}{|V|}, \frac{|V_{\text{query}}|}{|V|}, \frac{|V_{\text{comment}}|}{|V|} \right], \tag{6}$$

where $|V_{stance}|$ is the count for each stance and $|V|$ is the total number of replies. This augmented feature representation adds three benefits to our rumor verification model. First, insight into credibility, since replies with negative stances (e.g., Deny) or critical engagement (e.g., Query) are often associated with false rumors, while supportive replies may reinforce credibility [29][12]. Next, cross-feature interaction, combined with hierarchical or temporal features, stance distribution amplifies the model's ability to learn patterns in rumor propagation [10]. Finally, signal amplification in sparse data, in threads with limited replies, stance distribution provides aggregate signals that are more robust than individual post features [11].

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

The final harmonic unified input feature combines the aforementioned components as:

$$F = [f, s, \mathbf{h}_{\text{input}}] \tag{7}$$

D. An attention-based Encoder

The encoder in this work is designed to process and integrate diverse features from conversational threads, including f, s, and $\mathbf{h}_{\text{input}}$. It employs specialized fully connected layers to encode each type of feature before combining them into a unified representation for subsequent processing by an attention layer.

1) Final Post Embedding Encoder: The input here is a concatenated vector of the source post embedding and stance-specific aggregated embeddings f from (3). A Fully-connected Feed-forward Layer (FFL) with ReLU activation and dropout regularization is used to process f as:

$$\mathbf{h}_{\text{post}} = \mathbf{ReLU}(\mathbf{W}_{\text{post}}\mathbf{f} + \mathbf{b}_{\text{post}}), \tag{8}$$

where $\mathbf{W}_{\text{post}}$ and $\mathbf{b}_{\text{post}}$ are learnable parameters. The output is a latent representation of stance embeddings with dimension H.

2) Stance Distribution Encoder: Likewise, here the input is a normalized stance distribution vector s from (6). We process s with an FFL, which transforms the structural vector into a latent representation:

$$\mathbf{h}_{\text{dist}} = \mathbf{ReLU}(\mathbf{W}_{\text{dist}}\mathbf{s} + \mathbf{b}_{\text{dist}}), \tag{9}$$

where $\mathbf{W}_{\text{dist}}$ and $\mathbf{b}_{\text{dist}}$ are trainable weights and $\mathbf{h}_{\text{dist}} \in \mathbb{R}^n$ is the encoded stance distribution output.

3) Final Hierarchical Level Encoder: This sub-encoder uses the structural information $\mathbf{h}_{\text{input}}$ encoded as one-hot or positional vectors. An FFL transforms the structural vector into a latent representation as:

$$\mathbf{h}_{\text{struct}} = \mathbf{ReLU}(\mathbf{W}_{\text{struct}}\mathbf{h}_{\text{input}} + \mathbf{b}_{\text{struct}}), \tag{10}$$

where $\mathbf{W}_{\text{struct}}$ and $\mathbf{b}_{\text{struct}}$ are learnable parameters ans $\mathbf{h}_{\text{struct}} \in \mathbb{R}^m$ encodes structural information.

4) Attention mechanism: Attention mechanisms enable models to focus on pertinent parts of input sequences dynamically, enhancing the capture of intricate dependencies and relationships [32, 33, 34]. We extend this aspect to model interactions among encoded features ($\mathbf{h}_{\text{post}}$, $\mathbf{h}_{\mathbf{dist}}$, $\mathbf{h}_{\text{struct}}$), leveraging a Multi-Head Attention (MHA) mechanism [35], which allows parallel attention computations to capture diverse relationships within the input feature space. Thus, it enhances feature fusion, by attending across multiple features, aiding the model to effectively integrate information from diverse sources, ensuring that stance semantics, hierarchical levels, and stance distributions are considered together for predicting rumor veracity. Hence, the three latent representations are concatenated along a sequence dimension to form the input for the MHA mechanism as:

$$\mathbf{H}_{\text{input}} = \mathbf{Concat}(\mathbf{h}_{\text{post}}, \mathbf{h}_{\text{dist}}, \mathbf{h}_{\text{struct}}), \tag{11}$$

where $\mathbf{H}_{\text{input}} \in \mathbb{R}^{n+m+H}$. For each attention head i, the scaled dot-product attention computes with ReLU activation function:

$$\text{head}_i = \text{ReLU}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i, \tag{12}$$

where Q_i, K_i, V_i are the query, key, and value matrices derived from input feature encodings through linear projections of $\mathbf{H}_{\text{input}}$ with the respective weight matrices as:

$$Q_i = \mathbf{H}_{\text{input}} W_i^Q, \quad K_i = \mathbf{H}_{\text{input}} W_i^K, \quad V_i = \mathbf{H}_{\text{input}} W_i^V \tag{13}$$

The outputs from all heads are concatenated and passed through a final linear transformation:

$$\text{MultiHead}(\mathbf{H}_{\text{input}}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_n) W^o, \tag{14}$$

where n represents number of attention heads and W^o is a weight matrix. The attended features from MHA are then aggregated using mean pooling to produce the final unified representation:

$$\mathbf{h}_{\text{final}} = \text{Dropout}\left(\frac{1}{n}\sum_{i=1}^{n} \text{MultiHead}(\mathbf{H}_{\text{input}})\right), \tag{15}$$

where $Dropout$ is used for regularization.

E. Decoder

The decoder aims to take the output ($\mathbf{h}_{\text{final}}$) from the MHA as input and makes the rumor veracity prediction. It consists of a single dense layer and a softmax activation function to compute class probabilities for rumor veracity prediction. The feed-forward layer is leveraged to map the high-dimensional attended representation $\mathbf{h}_{\text{final}}$ to the output space corresponding to the rumor classes (True, False, Unverified):

$$\mathbf{z} = \text{Dropout}\left(\mathbf{W}_{\text{out}}\mathbf{h}_{\text{final}} + \mathbf{b}_{\text{out}}\right), \tag{16}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{C \times D}$ is a learnable weight matrix that maps the hidden representation $\mathbf{h}_{\text{final}}$ to the output classes C, while $\mathbf{b}_{\text{out}} \in \mathbb{R}^C$ is the bias term. In this case, D is the dimensionality of $\mathbf{h}_{\text{final}}$ and C = 3. The raw output $\mathbf{z}$ is passed through a softmax function to convert it into probabilities for each class:

$$\hat{y}_i = \text{softmax}(\mathbf{z_i}) = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}, \quad i = 1, \ldots, C, \tag{17}$$

where $\hat{y}_i$ is the predicted probability for class i.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

F. Objective function

During training, the predicted probabilities $\hat{y}$ are compared with the ground truth labels using the cross-entropy loss function:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{18}$$

where y_i is the one-hot encoded ground truth label.

The objective function for the rumor verification task is designed to minimize the classification error $\mathcal{L}$. That is, it integrates the outputs from the attention mechanism, processed by the decoder, and compares the predictions with the ground truth labels. In detail, the objective function is expressed as

$$\mathcal{J}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c}, \tag{19}$$

where N is the total number of rumor events in a training batch.

IV. Experiments

This section assesses the performance of our model in comparison to SOTA baselines and conducts a comprehensive analysis to gain deeper insights into the model's effectiveness.

A. Datasets

Experiments are conducted on three widely used and publicly available challenging benchmark datasets: SemEval-2017 [14], RumorEval-2019 [36], and PHEME [37]. Among these, RumorEval-2019 and PHEME extend the SemEval-2017 task, which comprises 325 rumor-related events and 5,568 tweets collected from eight major breaking news events.

On the one hand, RumorEval-2019 extends SemEval-2017 by incorporating additional test data and new Reddit-based content while utilizing all SemEval-2017 rumor events for training. It consists of 446 rumor-related conversational threads and a total of 8,574 posts. The claims in both SemEval-2017 and RumorEval-2019 are annotated with three veracity labels: True, False, or Unverified. Each post within a thread is assigned a stance label: Support, Deny, Query, or Comment. On the other hand, PHEME enhances RumorEval-2017 by incorporating additional rumor events and data from nine major breaking news stories on Twitter. It contains 2,402 conversational threads and 105,354 tweets. Unlike RumorEval-2019, the additional data in PHEME is annotated solely with rumor veracity labels.

These datasets are widely recognized within the research community. Employing the same datasets as previous studies enables a fair and direct comparison between our approach and SOTA techniques, ensuring consistency and reliability in the experimental evaluation.

Tables I, II, and III provide detailed statistics of the datasets. In the tables, (F, T, U) represent (False, True, Unverified), (S, D, Q, C) represent (Support, Deny, Query, Comment), and NS-M stands for Non-Stance Comments, respectively.

B. Data Preprocessing

Alongside conventional data preprocessing techniques, like removing null entries, we implement hashtag processing and text normalization, adapting the approach outlined by [38]. We additionally substitute all hyperlinks in the text with url and replace all *@user* mentions with $mention$, as these transformations demonstrated effectiveness in prior work [9].

C. Experimental setup

Our model utilizes the uncased BERT base [2] to generate word embeddings for both a claim c and replies R within a rumor event e_i. We also tested alternative pre-trained language models (PLMs), such as the Robustly Optimized BERT Approach (RoBERTa) [39] and the Generative Pre-trained Transformer (GPT) [40]. However, they yielded suboptimal performance compared to BERT and were therefore dropped in most of our experiments. During training, the model processes 16 rumor events per batch. The BERT tokenizer is configured with a maximum sequence length of 128. Optimization is performed using the Adam optimizer [41], with a learning rate of 0.0001. Other hyperparameters include a dropout probability of 0.35 and four attention heads. For encoding hierarchical levels and stance distributions, the embedding dimensions are set to [18, 19, 20] for [SemEval-2017, PHEME, RumorEval-2019], corresponding to the average thread lengths in these datasets. We also experimented with various dimensions to find optimal ones.

For SemEval-2017 and RumorEval-2019, we adhere to the standard train/validation/test split as defined in the original publications. Conversely, since PHEME does not provide an official dataset split, a conventional evaluation protocol is adopted, that follows a leave-one-out k-fold validation strategy, where each event is used as a test set in turn.

To address class imbalance, class weights are dynamically calculated based on label frequencies and incorporated into the cross-entropy loss function $\mathcal{L}$ during training. The model's performance is evaluated using the Macro-F1 score and accuracy metrics, with the best-performing model on the validation Macro-F1 score saved for final testing. All hyperparameters were meticulously fine-tuned using the development dataset. The reported results are aggregated from ten experimental runs. All experiments were conducted on two Quadro RTX 8000 GPUs, each equipped with 48 GB of VRAM, ensuring sufficient computational resources for our tasks.

As the PHEME dataset includes only partial stance annotations, we initially trained the model (omitting the stance-based embedding aggregation and stance distribution modules during this stage) using both the stance-labeled RumorEval-2019 and SemEval-2017 datasets. Given that these datasets exhibit a significant skew toward the *Comment* stance, we applied the SMOTE [42] oversampling technique to balance the stance distribution and improve the generalization of the model. The best-performing model from this training was then used to predict the stance labels for the PHEME dataset. On the same note, SMOTE was not used for rumor verification to ensure a fair comparison with baseline methods.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I
DETAILED STATISTICS OF RUMEVAL2017

Split	Rumor Statistics					Stance Distribution			
	#Threads	**AvgDepth**	**#F**	**#T**	**#U**	**#S**	**#D**	**#Q**	**#C**
Train set	272	3.2	50	127	95	841	333	330	2734
Development set	25	3.4	12	10	3	69	11	28	173
Test set	28	2.8	12	8	8	94	71	106	778
Total	**325**	**3.2**	**74**	**145**	**106**	**1004**	**415**	**464**	**3685**

TABLE II
DETAILED STATISTICS OF RUMEVAL2019

Split	Rumor Statistics					Stance Distribution			
	#Threads	**AvgDepth**	**#F**	**#T**	**#U**	**#S**	**#D**	**#Q**	**#C**
Twitter Train	325	2.2	74	145	106	1004	415	464	3685
Reddit Train	40	3.0	24	9	7	23	45	51	1025
Total Train	365	2.3	98	154	113	1027	460	515	4700
Twitter Test	56	0.9	30	22	4	141	92	62	771
Reddit Test	25	2.9	10	9	6	16	54	31	705
Total Test	81	1.7	40	31	10	157	146	93	1476
Total	**446**	**2.3**	**138**	**185**	**123**	**1184**	**606**	**608**	**6176**

TABLE III
DETAILED STATISTICS OF PHEME

Event	Rumor Statistics					Stance Distribution				
	#Threads	**AvgDepth**	**#F**	**#T**	**#U**	**#S**	**#D**	**#Q**	**#C**	**#NS-Com**
Charlie Hebdo	458	3.5	116	193	149	248	60	61	795	380
Sydney siege	522	3.3	86	382	54	225	90	110	769	448
Ferguson	284	5.2	8	10	266	191	95	116	784	234
Ottawa shooting	470	2.9	72	329	69	171	78	83	568	407
Germanwings-crash	238	3.1	111	94	33	80	16	43	244	209
Putin missing	126	2.2	9	0	117	18	6	5	33	117
Prince Toronto	229	2.3	222	0	7	21	7	11	64	217
Gurlitt	61	1.3	0	59	2	0	0	0	0	61
Ebola Essien	14	2.4	14	0	0	6	6	1	21	12
Total	**2402**	**3.6**	**638**	**1067**	**697**	**960**	**358**	**430**	**3278**	**2085**

D. Baseline Models

We present our model in three variations (CoSDD-Depth, CoSDD-Temp, and CoSDD-Breadth) each differentiated by the encoding strategy used for hierarchical levels. CoSDD-Depth and CoSDD-Breadth adopt DFS and BFS strategies, respectively, while CoSDD-Temp utilizes temporal sequencing. These variations are evaluated against several SOTA rumor detection models:

1) **Stance-Conditioned Modeling (S-CoM)** [1]: This is our prior work, that is extended in this paper. It aggregates post embeddings and separately models stance progression with a BiLSTM for rumor verification.
2) **eventAI** [43]: This method, which secured the first position in the RumorEval-2019 competition task [44], leverages multidimensional information and employs an ensemble strategy to enhance rumor verification.
3) **MTL2-Hierarchical Transformer** [8]: This approach segments conversational threads into multiple groups based on the hierarchical structure of conversations. Each group is processed using BERT to extract contextual information, and the aggregated information is fused using a Transformer for rumor verification.
4) **Coupled Hierarchical Transformer (CHT)** [8]: Building on the MTL2-Hierarchical Transformer, this model incorporates BERT to capture contextual nuances. It further enhances performance by integrating stance information.
5) **SEMTEC** [45]: This method introduces SEMTEC, a deep learning-based approach that combines emotion features, sentiment attributes, and contextual text analysis to improve rumor detection.
6) **Joint Rumor and Stance Model (JRSM)** [9]: This framework utilizes a graph transformer to encode input data and a partition filter network to model explicitly rumor-specific, stance-specific, and shared interactive features. These features are subsequently employed for

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

joint rumor and stance classification.

7) **SAMGAT** [31]: This model utilizes Graph Attention Networks (GATs) to capture contextual relationships between posts. While initially applied to the PHEME dataset for binary classification (excluding the *Unverified* class), we adapt and retrain SEMTEC and SAMGAT for our experimental setup (three classes) across all relevant datasets.
8) **Knowledge Graphs (KGs)** [46]: This study proposes a knowledge graph-based methodology that automatically retrieves evidence for rumor verification.

E. Results

1) Results and Discussion: Table IV provides a comparative analysis of the performance of the models. The findings demonstrate that our model significantly outperforms the best-competing baselines, as validated by McNemar's test with a p-value < 0.05 [47], with the *CoSDD-Breadth* variant delivering the best results. Furthermore, our results exhibit a standard deviation in the range of 0.006–0.02 across all three datasets over the 10 experimental runs, indicating robust and consistent performance.

From the table, while *eventAI* uses multidimensional information and ensemble learning to boost performance, it relies heavily on pre-defined multidimensional features, which could miss latent cues in data, and it does not account for the hierarchical structure of conversations or the distribution of stances, which are critical for understanding rumor propagation dynamics. *MTL2-Hierarchical Transformer* processes hierarchical conversational threads by segmenting them into groups and using BERT to extract contextual features. It further employs a Transformer to fuse group-level information. Still, the hierarchical structure is only partially encoded, and inter-group interactions may be underrepresented and does not leverage stance distributions as part of its input for understanding conversations in rumor detection and also fails to emphasize temporal dependencies, which can provide critical context for rumor dynamics. *Coupled Hierarchical Transformer* introduces an attention mechanism to integrate stance information, which adds contextual nuance but the stance information is not as comprehensive or dynamically aligned with structural and temporal features as in our model, the model lacks a systematic representation of hierarchical dependencies.

The *Joint Rumor and Stance* baseline model effectively combines rumor and stance-specific features, but it also lacks the temporal and structural encoding capabilities that our model achieves via *CoSDD-Breadth* and *CoSDD-Temp*. *SAMGAT's* reliance on localized graph attention limits its ability to capture global relationships across threads, which our model handles effectively through hierarchical encodings. Compared to SAMGAT and SEMTEC, our model is more adaptable to three-class classification, as demonstrated by the substantial performance boost, since these two baselines were primarily intended for binary rumor classification.

The evidence-retrieval approach of *Knowledge Graphs* is limited by the availability and quality of external evidence. S-Com is close to our model's performance, but it statically models stance progression with BiLSTM without accounting for dynamic stance sequencing strategies in a conversational tree. Our models leverage hierarchical encoding to capture the complex structural relationships in conversational threads effectively. Including stance distributions as part of the input features, concurrently applying graph traversal techniques to hierarchical encoding, enables the model to align stance dynamics with rumor classification. This aligns with methods like the *Joint Rumor and Stance* model but incorporates richer attention mechanisms, providing a significant edge. The MHA mechanism ensures that the model focuses on relevant parts of the input (e.g., stance signals and structural levels). Our models enhance this by modeling the entire structural hierarchy effectively, compared to models like *SAMGAT* that focus on localized graph attention. Our approach incorporates embeddings aggregated from source posts, stance distributions, and structural hierarchies, allowing for nuanced feature representations. This approach captures latent cues that other baselines (e.g., *eventAI* and *SEMTEC*) might overlook due to reliance on pre-determined features like sentiment or emotion.

It has been noticed from the table that the *CoSDD-Breadth* variant of our model performs better than the *CoSDD-Depth* and *CoSDD-Temp* variants, and we point out the following factors. To begin with, *CoSDD-Breadth* processes nodes level by level within the hierarchical structure of a conversation thread. This allows the model to capture the broad, overall structure and relationships across different levels of the hierarchy simultaneously. In contrast, *CoSDD-Depth* focuses on deeper, more localized paths first, which might cause the *CoSDD-Depth* model to miss or underrepresent cross-level interactions that are critical for understanding the broader context of rumor propagation. Secondly, rumors often involve interactions across multiple users at the same hierarchical level (e.g., multiple replies to a single source post or other replies). *CoSDD-Breadth* captures these parallel interactions effectively by processing all posts at the same level together. *CoSDD-Depth*, on the other hand, focuses on specific threads or chains of replies, which may lead to a narrow perspective on the overall conversational dynamics and off-topic diversions. Third, *CoSDD-Breadth* ensures that stance distributions across the hierarchical levels are aggregated and attended to in a balanced way, as the model processes the entire level before moving to the next. This aligns well with our model's attention mechanism, which uses these distributions to refine the representation of conversational threads. The *CoSDD-Depth* variant, conversely, may prioritize deeper paths without adequately integrating the stance dynamics across intermediate hierarchical levels. Finally, the *CoSDD-Temp* variant primarily relies on time-ordered posts, which may not always align with the hierarchical structure. In cases where multiple posts are created simultaneously or out of sequence, temporal encoding can misrepresent the structural relationships. However, the *CoSDD-Breadth* naturally respects the hierarchical structure

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE IV
PERFORMANCE COMPARISON WITH BASELINE MODELS.

Model	SemEval-2017		RumorEval-2019		PHEME	
	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc
eventAI	0.618	0.629	0.577	0.591	0.342	0.357
MTL2-Hierarchical Transformer	0.657	0.643	0.568	0.572	0.375	0.454
Coupled Hierarchical Transformer	0.680	0.678	0.579	0.611	0.396	0.466
SAMGAT	0.702	0.709	0.542	0.562	0.409	0.418
SEMTEC	0.711	0.727	0.581	0.592	0.421	0.437
Joint Rumor and Stance Model	0.754	0.767	0.598	0.623	0.448	0.479
Knowledge Graphs	0.758	0.759	0.584	0.593	0.489	0.523
S-Com	0.774	0.781	0.636	0.648	0.641	0.643
CoSDD-Depth	0.775	0.767	0.712	0.724	0.591	0.622
CoSDD-Temp	0.776	0.789	0.728	0.731	0.595	**0.659**
CoSDD-Breadth	**0.783**	**0.798**	**0.731**	**0.765**	**0.658**	0.638

while maintaining a balance between temporal and structural features. This makes it more robust for predicting rumors from online discourse, where hierarchical relationships are crucial.

Although only Twitter and Reddit data are used in our experiments, this work can be customized and extended to any social media platform actively engaging in fact-checking and where users participate in the subsequent conversations about a source claim. Therefore, our stance-conditioned modeling for rumor verification can also be generalized to Facebook, Instagram, Threads, etc. This will be incorporated into future work.

TABLE V
ABLATION STUDY OF OUR MODEL ON THE RUMOREVAL-2019 AND PHEME DATASETS.

Model	RumorEval-2019		PHEME	
	Macro-F1	Acc	Macro-F1	Acc
-structs & emb aggreg	0.540	0.566	0.345	0.582
-structs	0.549	0.573	0.417	0.590
-emb aggreg	0.552	0.579	0.410	0.623
-hier levels	0.561	0.582	0.450	0.593
-stance distr	0.573	0.594	0.461	0.628
-MHA	0.595	0.642	0.474	0.631
CoSDD-Breadth	**0.731**	**0.765**	**0.658**	**0.638**

F. Ablation Study

The ablation study, as illustrated in Table V, provides insights into the contributions of different modules within our model. By systematically removing individual components and observing the changes in Macro-F1 and accuracy on the RumorEval-2019 and PHEME datasets, we can evaluate the importance of each module. In the table, *-structs & emb aggreg* indicates that our model relies solely on the claim post c, omitting both the structural dynamics from replies R and the embeddings aggregation $\mathbf{e}_{\text{stance}}$. The *-structs* configuration excludes the stance distribution $\mathbf{h}_{\text{dist}}$ and the hierarchical levels encoding $\mathbf{h}_{\text{struct}}$. In *-emb aggreg*, the embeddings aggregation mechanism is removed; instead, the entire rumor event is encoded as a single BERT embedding, constrained by the maximum sequence length of the language model (i.e., 512). The variant *-hier levels* excludes the hierarchical levels encoding feature, while *-stance distr* removes only the stance distribution module. The *-MHA* configuration does not utilize the MHA mechanism; instead, it directly employs the combined feature vector F for predicting rumor veracity. Lastly, *CoSDD-Breadth* represents our most comprehensive model, incorporating all modules and employing BFS traversal to encode hierarchical levels. Here is an analysis of the results:

1) Impact of Structural Dynamics and Embedding Aggregation (-structs & emb aggreg): Removing both structural dynamics and embedding aggregation mechanisms results in the worst performance across both datasets (Macro-F1 drops to 0.540 on RumorEval-2019 and 0.345 on PHEME). The model, in this configuration, relies solely on the claim post c, completely ignoring the hierarchical structure and stance information from the replies R. Additionally, excluding the embedding aggregation only, meanwhile encoding the entire rumor event as a single BERT embedding (*-emb aggreg*), further limits the model's ability to distinguish nuanced relationships between posts. This confirms that capturing relationships within conversation threads and integrating stance distributions is essential for effective rumor verification.

2) Impact of Structural Dynamics (-structs, -hier levels, -stance distr): Excluding both/either stance distribution and/or hierarchical level encodings results in a significant performance drop compared to the full model. The hierarchical structure of a conversation plays a critical role in understanding rumor propagation. Without this structural context, the model cannot effectively capture the interplay between a claim post c and replies R, and the model cannot represent how information propagates through different levels of the conversation. The stance distribution is crucial for incorporating the stance dynamics of replies into the rumor veracity decision. Without it, the model loses critical features that differentiate between supportive, commenting, denying, and querying stances. This highlights the importance of structural dynamics ($\mathbf{h}_{\text{struct}}$ and $\mathbf{h}_{\text{dist}}$) in improving veracity prediction.

3) Impact of Multi-Head Attention (-MHA): Removing the MHA mechanism equally reduces the performance of the model. MHA enhances the model by allowing it to focus on different aspects of the conversation, such as structural relationships, stance, and temporal cues. Directly using the

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

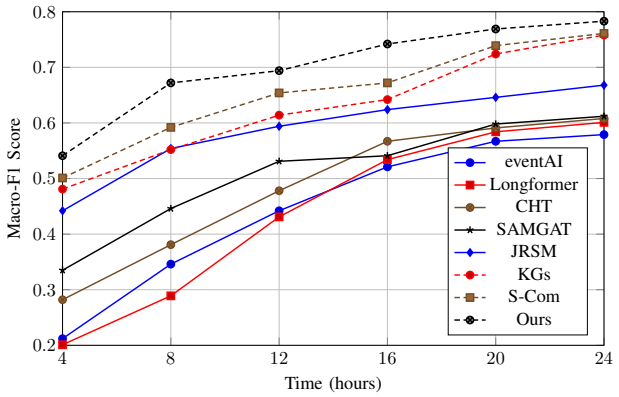


Fig. 3. Early Rumor Detection Performance on SemEval-2017.

combined feature vector reduces the model's flexibility and representational power.

The full model (*CoSDD-Breadth*) achieves the highest performance on both datasets, with a Macro-F1 of 0.731 on RumorEval-2019 and 0.658 on PHEME. Combining all modules ensures that the model captures both the global and local dynamics of a discourse. The BFS traversal specifically enhances the representation by effectively encoding hierarchical relationships across levels.

G. Early Rumor Detection

Timely detection of rumors can mitigate their widespread dissemination. To assess early rumor detection capabilities, we define detection checkpoints based on the elapsed time, spanning 24 hours, since the initial post. At each checkpoint, only replies accumulated up to that point are considered for model evaluation, and performance is measured using the Macro-F1 score at each detection interval.

Figure 3 illustrates Macro-F1 and accuracy scores over time for early rumor detection on the SemEval-2017 dataset. Our model consistently outperforms all baselines throughout the 24-hour period, demonstrating superior effectiveness in detecting rumors early. While all models improve as more information becomes available, our model achieves significantly higher Macro-F1 scores early on, starting with an advantage at 4 hours and maintaining superior performance throughout. This suggests that our approach is more responsive to limited initial data, making it highly effective for early-stage rumor identification and particularly valuable in real-world misinformation scenarios where timely intervention is crucial.

H. Performance of different pre-trained language models

To identify the most effective pre-trained language model for generating contextual embeddings of rumor events, we conducted additional experiments comparing BERT, RoBERTa, and GPT. These models were chosen due to their strong presence and performance in the existing literature. For each language model, we applied all three configurations of our approach: *Depth*, *Temporal*, and *Breadth*. The performance results for each configuration across the SemEval-2017, RumorEval-2019, and PHEME datasets are summarized in Table VI.

The results indicate a clear performance advantage for BERT-based configurations, which consistently outperform both GPT and RoBERTa counterparts across all datasets and metrics. Among the three configurations, CoSDD-Breadth-BERT achieves the highest overall scores on the SemEval-2017 dataset, with a Macro-F1 of 0.783 and an accuracy of 0.798, while also attaining top performance on RumorEval-2019 (Macro-F1 = 0.731, Acc = 0.765). On the PHEME dataset, BERT models also dominate, with CoSDD-Temp-BERT achieving the best accuracy (0.659) and CoSDD-Breadth-BERT obtaining the highest Macro-F1 (0.658). In contrast, GPT-based models show relatively lower performance, with Macro-F1 scores generally below 0.73 for SemEval-2017 and below 0.70 for RumorEval-2019. RoBERTa models perform better than GPT but still fall short of BERT, suggesting that BERT's embedding representations are more effective for this rumor verification task, particularly when combined with the *Breadth* configuration.

V. Conclusion

This study presents a novel approach to rumor verification that leverages hierarchical structural encoding, stance distributions, and MHA mechanisms to capture the intricate dynamics of rumor propagation in conversational threads. Our model demonstrates SOTA performance across multiple benchmark datasets, including SemEval-2017, RumorEval-2019, and PHEME, outperforming existing approaches with statistically significant improvements in Macro-F1 and accuracy metrics. Our findings highlight several key insights: the explicit incorporation of stance information significantly improves rumor verification, demonstrating that user reactions provide crucial contextual cues. Our experiments further show that breadth-based graph traversal outperforms depth-based and temporal-based sequencing strategies for hierarchical encodings. Early rumor detection analysis demonstrates that our model achieves faster and more accurate misinformation detection than competing methods, underscoring its practical utility in real-world misinformation detection; meanwhile, BERT indicated stronger performance over RoBERTa and GPT when encoding discourse embeddings.

While the model has shown success, its limitations include a heavy reliance on accurate stance annotations—which might not be consistently available—and training on datasets that may not fully represent real-world misinformation trends across diverse social media platforms. Additionally, the focus on textual content ignores the visual aspects (such as images, memes, and videos) that often accompany online rumors. Future work could reduce dependence on manually labeled data through weakly supervised and self-supervised learning, improve generalization via cross-platform adaptation, incorporate multi-modal data, and further explore extra structural

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE VI
PERFORMANCE COMPARISON OF PRE-TRAINED LANGUAGE MODELS.

Model	SemEval-2017		RumorEval-2019		PHEME	
	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc
CosDD-Depth-GPT	0.709	0.719	0.682	0.687	0.522	0.551
CosDD-Temp-GPT	0.712	0.723	0.689	0.691	0.523	0.536
CosDD-Breadth-GPT	0.725	0.731	0.692	0.703	0.541	0.545
CosDD-Temp-RoBERTa	0.721	0.739	0.698	0.713	0.558	0.575
CosDD-Depth-RoBERTa	0.734	0.742	0.706	0.710	0.545	0.552
CosDD-Breadth-RoBERTa	0.748	0.761	0.712	0.715	0.557	0.601
CoSDD-Depth-BERT	0.775	0.767	0.712	0.724	0.591	0.622
CoSDD-Temp-BERT	0.776	0.789	0.728	0.731	0.595	**0.659**
CoSDD-Breadth-BERT	**0.783**	**0.798**	**0.731**	**0.765**	**0.658**	0.638

dynamics like stance distribution, hierarchical level encoding, and attention mechanisms.

Finally, while the work has positive implications, ethical challenges and risks persist. False negatives and false positives could respectively suppress credible information or allow misinformation to spread, so human validation of predictions is recommended. The system's success could also enable misuse, such as censorship or targeting, requiring transparent deployment and strict ethical guidelines. Additionally, training data biases may lead to unfair outcomes; therefore, evaluating and mitigating these biases is crucial.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) under Award number OIA-1946391, Data Analytics that are Robust and Trusted (DART).

REFERENCES

[1] G. Nkhata and S. Gauch, "Stance-conditioned modeling for rumor verification," in *The Seventeenth International Conference on Information, Process, and Knowledge Management (eKNOW 2025)*. Nice, France: IARIA, May 2025, pp. 30–36, copyright © IARIA, 2025.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[3] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[4] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aap9559

[5] Y. Özkent, "Social media usage to share information in communication journals: An analysis of social media activity and article citations," *PLOS One*, vol. 17, no. 2, p. e0263725, 2022.

[6] E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task learning for rumour verification," *arXiv preprint arXiv:1806.03713*, 2018.

[7] J. Manurung, P. Sihombing, M. A. Budiman *et al.*, "Dynamic rumor control in social networks using temporal graph neural networks," in *2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*. IEEE, 2023, pp. 1–5.

[8] J. Yu, J. Jiang, L. M. S. Khoo, H. L. Chieu, and R. Xia, "Coupled hierarchical transformer for stance-aware rumor verification in social media conversations." Association for Computational Linguistics, 2020.

[9] N. Luo, D. Xie, Y. Mo, F. Li, C. Teng, and D. Ji, "Joint rumour and stance identification based on semantic and structural information in social networks," *Applied Intelligence*, vol. 54, no. 1, pp. 264–282, 2024.

[10] H. Gong, M. Zhang, Q. Liu, S. Wu, and L. Wang, "Breaking event rumor detection via stance-separated multi-agent debate," *arXiv preprint arXiv:2412.04859*, 2024.

[11] J. Bian *et al.*, "Multi-task learning for stance and rumor detection in online conversations," in *Proceedings of ICWSM*, 2023.

[12] Y. Bai, C. Han, and Y. Jia, "Samgat: Structure-aware multilevel graph attention networks for automatic rumor detection," *PeerJ Computer Science*, vol. 9, p. e1032, 2024.

[13] P. Yang, J. Leng, G. Zhao, W. Li, and H. Fang, "Rumor detection driven by graph attention capsule network on dynamic propagation structures," *The Journal of Supercomputing*, vol. 79, pp. 15 674–15 690, 2023.

[14] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours," *arXiv preprint arXiv:1704.05972*, 2017.

[15] J. Li, Y. Bin, Y. Ma, Y. Yang, Z. Huang, and T.-S. Chua, "Filter-based stance network for rumor verification," *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1–28, 2024.

[16] E. Elmurngi and A. Gherbi, "An empirical study on detecting fake reviews using machine learning techniques," in *2017 Seventh International Conference on Innovative Computing Technology (INTECH)*. IEEE, 2017, pp. 107–114.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[17] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2018.

[18] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "Ti-cnn: Convolutional neural networks for fake news detection," *arXiv preprint arXiv:1806.00749*, 2018.

[19] Q. Lv, Y. Wang, B. Zhang, and Q. Jin, "Rv-ml: An effective rumor verification scheme based on multi-task learning model," *IEEE Communications Letters*, vol. 24, no. 11, pp. 2527–2531, 2020.

[20] Y. Luo, J. Ma, and C. K. Yeo, "Bcmm: A novel post-based augmentation representation for early rumour detection on social media," *Pattern Recognition*, vol. 113, p. 107818, 2021.

[21] Y. Wang, B. Zhang, J. Ma, and Q. Jin, "Marv: Multi-task learning and attention based rumor verification scheme for social media," in *2022 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2022, pp. 94–98.

[22] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 1980–1989.

[23] C. Yuan, Q. Ma, W. Zhou, J. Han, and S. Hu, "Jointly embedding the local and global relations of heterogeneous graph for rumor detection," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 796–805.

[24] Y.-J. Lu and C.-T. Li, "Gcan: Graph-aware co-attention networks for explainable fake news detection on social media," *arXiv preprint arXiv:2004.11648*, 2020.

[25] L. Wei, D. Hu, W. Zhou, Z. Yue, and S. Hu, "Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional networks for rumor detection," *arXiv preprint arXiv:2107.11934*, 2021.

[26] Q. Mai, S. Gauch, D. Adams, and M. Huang, "Sequence graph network for online debate analysis," 2025. [Online]. Available: https://arxiv.org/abs/2406.18696

[27] H. Sahni *et al.*, "Temporal dynamics in rumor propagation: Challenges and solutions," in *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 2022.

[28] J. Choi, T. Ko, Y. Choi, H. Byun, and C.-k. Kim, "Dynamic graph convolutional networks with attention mechanism for rumor detection on social media," *PLOS One*, vol. 16, no. 8, p. e0256039, 2021.

[29] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PLOS One*, vol. 11, no. 3, p. e0150989, 2016.

[30] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the EMNLP-IJCNLP 2019*, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[31] Y. Li, Z. Chu, C. Jia, and B. Zu, "Samgat: structure-aware multilevel graph attention networks for automatic rumor detection," *PeerJ Computer Science*, vol. 10, p. e2200, 2024.

[32] X. Li, Y. Zhou, and J. Liu, "Weighted graphsage with attention mechanism for rumor detection," *Scientific Reports*, vol. 13, no. 1, p. 12345, 2023.

[33] H. Wang, L. Zhao, and K. Liu, "Multi-attention neural interaction networks for rumor detection," in *Proceedings of the IEEE International Conference on Big Data*. IEEE, 2023, pp. 1023–1034.

[34] W. Zhang, Y. Chen, and Z. Li, "Attention graph adversarial dual contrast learning for rumor detection in social media," *PLOS One*, vol. 18, no. 8, p. e0290291, 2023.

[35] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[36] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski, "SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 845–854. [Online]. Available: https://aclanthology.org/S19-2147

[37] A. Zubiaga, M. Liakata, and R. Procter, "Learning reporting dynamics during breaking news for rumour detection in social media," *arXiv preprint arXiv:1610.07363*, 2016.

[38] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a# twitter," in *Proceedings of the 49th Annual meeting of The Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 368–378.

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," in *arXiv preprint arXiv:1907.11692*, 2019.

[40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[41] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[43] Q. Li, Q. Zhang, and L. Si, "eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds. Minneapolis, Minnesota, USA: Association for

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Computational Linguistics, Jun. 2019, pp. 855–859. [Online]. Available: https://aclanthology.org/S19-2148

[44] G. Gorrell, K. Bontcheva, L. Derczynski, E. Kochkina, M. Liakata, and A. Zubiaga, "Rumoureval 2019: Determining rumour veracity and support for rumours," *arXiv preprint arXiv:1809.06683*, 2018.

[45] D. Sharma and A. Srivastava, "Detecting rumors in social media using emotion based deep learning approach," *PeerJ Computer Science*, vol. 10, p. e2202, 2024.

[46] J. Dougrez-Lewis, E. Kochkina, M. Liakata, and Y. He, "Knowledge graphs for real-world rumour verification," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 9843–9853.

[47] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

People First, Platform Second: Designing Hybrid Learning for Emotional and Educational Resilience in Unsafe Regions

Maria Eisenmann
English Teaching Methodology Chair
Julius-Maximilians-Universität Würzburg
Würzburg, Germany
e-mail: maria.eisenmann@uni-wuerzburg.de

Nataliia Lazebna
Centre for Teaching and Learning
Julius-Maximilians-Universität Würzburg
Würzburg, Germany
e-mail: nataliia.lazebna@uni-wuerzburg.de

Kateryna Lut
Foreign Philology and Translation Department
National University "Zaporizhzhia Polytechnic"
Zaporizhzhia, Ukraine
e-mail: katerinalut@gmail.com

***Abstract*—In regions where education is disrupted by conflict and instability, hybrid learning models have emerged as critical tools for maintaining access to learning and emotional support. This study explores the technological, anthropocentric, and socio-emotional challenges and solutions associated with hybrid-format learning in an unsafe region (Ukraine). The paper is based on the teaching experiences gained during hybrid seminars on (1) Teaching English as a Foreign Language (2022/2023) for student teachers from Germany's Julius-Maximilians-Universität Würzburg (JMU) and Ukraine's National University "Zaporizhzhia Polytechnic" (NUZP), and hybrid seminars on (2) Transcultural Project-Based Learning. Multilingualism through the Arts (Summer Semester, 2025) for students from JMU, Ukraine's Vasyl Stus Donetsk National University, and National University "Zaporizhzhia Polytechnic". Technological limitations, well-being challenges, available solutions, and perspectives for further development of "hybrid harbors" based on Activity-Centered Analysis and Design are considered. The paper presents practical solutions to the challenges of foreign language teaching in conflict-affected contexts.**

Keywords-hybrid; physical, epistemic, and social aspects of learning; cozy corners; tech pods; empathy and resilience, unsafe regions.

I. INTRODUCTION

In the evolving landscape of education, Hybrid-Format Learning (HFL) emerges as a postdigital learning space, blending digital and physical environments. This study examines the technological, anthropocentric, and socio-emotional challenges, as well as potential solutions, associated with HFL in an unsafe region, specifically Ukraine. The objective of this research is to determine how hybrid learning can contribute to the development of social-emotional competencies and students' mental resilience. Participants in HFL seminars on (1) Teaching English as a Foreign Language (TEFL) (2022/2023), future EFL student teachers from Germany's Julius-Maximilians-Universität Würzburg and Ukraine's National University "Zaporizhzhia Polytechnic", as well as (2) Transcultural Project-Based Learning. Multilingualism through the Arts (Summer Semester, 2025) involving JMU, Ukraine's National University "Zaporizhzhia Polytechnic" (1), and Vasyl Stus Donetsk National University (2). The Ukrainian University (1) is situated close to the front line and periodically suffers from shelling that damages its buildings and makes classroom learning impossible. Some students and teachers have fled the region or the country to find safer living conditions. Those who stayed continue to experience electricity disruptions due to ongoing military conflict actions in the country. The Ukrainian University (2) is a University in exile, which was displaced to Vinnytsia city in the western part of Ukraine in 2014 after the occupation of Donetsk city by Russian forces, the original location of the university. The only way to continue education and maintain a high level of students' engagement in the current situation is to carry out the learning process in a digital environment.

The necessity of a safe digital learning environment during crises, such as the ongoing war in Ukraine, is indisputable. Working together in an HFL environment, urgently designed with the basic technologies available (laptops and a projector) during the first series of seminars (1), German and Ukrainian students were able to foster their

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

professional competencies, emotional intelligence, empathy, and resilience. To avoid "a feeling of loneliness in a hybrid-format environment that can be as harmful as smoking one box of cigarettes a day" [2], their teachers fostered trust and safety by bringing all voices into the room through a network of interrelated topics and dialogues. The second seminar was held in the facilities of the hybrid classroom at JMU [3], which were equipped with the necessary technologies.

The paper provides a more detailed investigation and expansion of the research initially outlined in the conference proceedings paper [1]. The current paper is structured as follows. In Section II, we present the background of this study. In Section III, we focus on the chosen methodology. In Section IV, we describe the main findings and discussion. Finally, we conclude in Section V and provide some future perspectives.

II. Background of the Study

The background information below provides insights into the development of hybrid learning, the challenges faced by teachers and students, and potential solutions that can be implemented. It also outlines the role of hybrid learning in supporting social-emotional development, fostering intercultural sensitivity, and enhancing mental resilience.

A. Hybrid Learning: Well-being and Social-Emotional Growth

Hybrid learning has emerged as a dominant and lasting approach in the post-pandemic educational landscape, moving beyond emergency remote teaching to integrated, innovative, and flexible designs. Lo (2023) notes that the "post-Covid-19 classroom is increasingly evolving towards a flipped, hybridised design that integrates students into the learning strategy and course structure through digital, participative, and adaptive settings" [4]. This evolution, however, has not been without its challenges, with Griffard et al. (2025) quoting an educator who described the initial phase as "This is not online education. It's disaster education" [5]. At the beginning of the pandemic, there were cases in Ukrainian schools when teachers opposed to having online classes, arguing that they were ineffective and merely a waste of time and effort.

A central theme across the relevant sources is the significant impact of these new learning environments on students' emotional well-being. Pipia & Williams (2024) highlight "heightened stress and anxiety fueled by uncertainties surrounding the post-pandemic" [6]. Conwi et al. (2024) quantitatively identified three mental health groups among students in blended learning: "flourishing (21.14%), languishing (39.72%), and moderately mentally healthy (39.14%)," with students generally displaying "moderate mental well-being, high academic distress, moderate familial distress, and strong adjustment to blended learning" [7]. Qualitative data from the same study revealed "emotional states, ranging from resilience and adaptability to anxiety and frustrations," alongside challenges such as "technological barriers, time management dilemmas, and exacerbation of isolation." Similar concerns about student isolation emerged in both fully remote and hybrid settings [5].

In conflict-affected regions, specifically Ukraine, these challenges are further amplified. Students and educators are not only exposed to technological limitations and pedagogical restructuring but also face existential threats and trauma resulting from war-related experiences. Under such challenging circumstances, mental health literacy becomes not just desirable but necessary for both teachers and students. The hybrid EFL classroom thus takes on a transformative role. Its unique combination of cultural, linguistic, and literary approaches offers a powerful space to raise awareness of mental health topics and destigmatise related discourse [8].

Socio-emotional learning (SEL) is consistently presented as a vital component for navigating the complexities of hybrid education. Petry et al. (2025) ground their Virtual Campus project in "the theoretical models of Self-Determination Theory (SDT, 1) and Social-Emotional Learning (SEL, 2)" [9]. Geesa et al. (2022) emphasize that "to facilitate o[nline]-SEL, educators need professional learning regarding content and ways to present SEL throughout online platforms" [10]. Moore (2021) asserts the necessity of "seamlessly integrat[ing] effective SEL practices into their online instruction, beginning with the all-important creation of a social learning community" [11].

There is a clear call for SEL to be viewed not as a "quick fix for problem behaviors or a crisis intervention," but as "a process, an approach practiced over time" [11]. The transition to hybrid models and the increased focus on emotional well-being have placed significant demands on educators. Sharma & Bhasin (2023) specifically address the "Well-being of faculty members in the era of hybrid education" [12]. Teachers also experience "burnout, being unsure who was listening in on class conversations, and feeling disconnected in an online environment" when implementing SEL during a crisis [13]. Issues with the Internet connection and sound disruptions can also add to the overall stress of remote teaching and learning, making it harder to concentrate on students' well-being. Challenges for instructors designing HyFlex courses include providing consistently "very similar classroom and learning experiences for both groups of students: remote and Face to Face" [14, 15]. The "New Normal" profoundly affects students' "emotional well-being, academic achievements, and overall performance" [6].

Positive emotional regulation and self-efficacy derived from well-designed hybrid experiences can "predict future attainment and learner competency" [4]. According to Conwi et al. (2024): "adjustment to the demands of blended learning can mitigate academic-related stressors" [7]. HyFlex models offer students "additional autonomy," which aligns with Self-Determination Theory and is theorized to increase students' sense of motivation for learning [14]. The Virtual Campus aims to foster intrinsic motivation by addressing basic psychological needs for autonomy, competence and social integration [9]. SEL encompasses "intrapersonal and interpersonal skills and development". Key SEL competencies include self-awareness, self-management,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

relationship skills, social awareness, and responsible decision-making.

In the context of Ukraine, the hybrid EFL classroom provides a unique space to engage with SEL through culturally responsive and trauma-sensitive pedagogy. English language education, particularly when linked to themes of identity, community, and interculturality, fosters the development of coping strategies, resilience, and interpersonal awareness. As outlined in Eisenmann et al. [8], it fosters mental health literacy by enabling students to recognize and articulate emotional distress, reflect on stigmas, and connect their personal experiences to global discourses. This is especially relevant for displaced or war-affected students, many of whom have experienced the disintegration of support networks.

Teachers have employed strategies such as "focusing on relationships; building routines and predictability; creating space to identify and share feelings; incorporating movement, mindfulness, and play; implementing culturally affirming practices; providing student choice and leadership; and engaging and collaborating with families" [15]. Meland & Brion-Meisels (2024) propose an integrative model for CS-SEL based on three core adult competencies: 1) "engaging in critical reflection"; 2) "building caring, authentic, and reciprocal relationships"; and 3) "shifting the balance of power toward the developing students". This model advocates for SEL as a "dynamic and reciprocal process between teacher and student," moving away where adults solely direct content. Co-construction and co-regulation are the central facilitative processes that are key to CS-SEL. Co-construction involves teachers and students as "agents in the co-creation of knowledge, skill development, and the classroom environment" [15]. This collaborative approach brings "valuable insights into a learning process."

The co-constructed hybrid classroom, as implemented in Ukrainian universities affected by war, enables learners not only to continue their education but also to develop emotional vocabulary, intercultural sensitivity, and mental resilience, which is essential in coping with both personal and collective crises [8].

There is a significant need for teacher training in designing online environments that foster a friendly and supportive atmosphere for students affected by trauma. Effective professional learning for o[nline]-SEL should be "relevant and readily transferable," offer "opportunities for interaction and feedback," require "active learning," allow "space for teacher reflection," and include "videos" [10]. It should also equip educators to be "equity leaders" and use "culturally responsive teaching." CASEL's SAFE acronym outlines four approaches: "Sequenced, Active, Focused, and Explicit" to effectively deliver SEL to all learners [10]. The specific nature of the virtual classroom is developed in a mutually beneficial process, which gives students space to know themselves and teachers time to know their students.

For further development of scenarios within the Activity-Centered Analysis and Design (ACAD) framework for emotional well-being in hybrid classrooms in unsafe regions, the following implications and recommendations based on the literature review will be considered:

1) Integrate Mental Health Education: Educational interventions should integrate mental health education into curricula in a holistic manner to foster well-being in an inclusive blended learning environment [7].

2) Design Emotionally Supportive Learning Experiences: The design and structure of hybrid learning must actively support positive emotional regulation [15]. This includes creating a social learning community [10].

3) Prioritize Educator Well-being and Training: Institutions must address the well-being of faculty members in hybrid education [12] and provide targeted professional development for educators on integrating SEL effectively into online and hybrid settings [9, 17].

4) Foster Student Voice and Agency: Empowering students to be in control of their learning accelerates progress [11]. Strategies like "Owning My Story, Goal Setting Postcard, and Service Learning Project" can facilitate this.

5) Cultivate Emotionally Intelligent Educators: Teachers and leaders with high socio-emotional competence are better prepared to engage students in SEL. This involves developing norms, cultivating relaxed alertness, and practicing mindfulness [11, 18].

6) Embrace Co-construction in SEL: Move towards a model where SEL is a dynamic and reciprocal process, with teachers and students co-creating learning environments and opportunities that meet diverse needs [16]. This also helps shift power dynamics in the classroom.

7) Utilize Technology to Enhance SEL: Platforms like the "Virtual Campus" can offer asynchronous learning, self-directed opportunities, and communicative/collaborative features that address psychological needs for autonomy, competence, and social integration, thereby strengthening SEL processes [9].

In conclusion, successfully navigating the "New Normal" in higher education necessitates a holistic approach that prioritizes student and educator well-being through intentional, culturally sustaining, and well-supported Social-Emotional Learning initiatives, harmoniously integrated into evolving hybrid learning models.

B. Technological limitations and solutions

Despite technological limitations, such as overloaded learning management systems, software constraints, student frustration with technology, sound failures, memory availability issues, and unstable internet connections, especially in an unsafe region [3-7], solutions at-hand were implemented (Seminar 1). At this point, the Miro whiteboard and/or Etherpad were used continuously. Additionally, despite the use of a single camera in the onsite classroom and the individual cameras of students on their cell phones, laptops, and other gadgets, missed discussions, visual cues, interactions, audio interruptions, and microphone issues hindered effective communication between online and face-to-face students. However, the students were provided with constant access to learning materials, group work, and discussion outputs were offered before and after HFL seminars.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

To address these challenges, the study recommends comprehensive training in technology and course management [10] as well as strategies to enhance the learner's experience and technology design [9-11]. Solutions include equipping classrooms with multiple cameras and high-quality microphones to capture all interactions and voices, thus fostering a more inclusive and dynamic learning environment. Ergonomic limitations also pose significant challenges, as lecturers often find themselves tethered to computers, reducing their mobility and engagement with students. The study suggests that both teachers and students should have the flexibility to move freely within the classroom to facilitate group work and engage in a more interactive learning process. However, technologies and equipment should be adapted accordingly, for example, to "catch" teachers' movements in the classroom or capture the faces and voices of the onsite session participants.

The inclusive, friendly, and flexible learning environment should host and nurture both students and teachers. For example, some teachers, invited speakers, and students have experienced either forced or voluntary displacement in their lives, and finding themselves together within a "hybrid harbor" helped them form closer connections rather than experience a feeling of alienation.

III. METHODOLOGY

The section outlines the methodological framework used in the study, focusing on Activity-Centered Analysis and Design and the development of scenarios for the hybrid classroom.

A. *Activity-Centered Analysis and Design*

The implementation of the Activity-Centered Analysis and Design (ACAD) framework fostered the physical, epistemic, and social aspects of learning. It was chosen as a practical guideline to solve challenging learning situations, namely "…Activity-Centered Analysis and Design (ACAD) is a meta-theoretical framework for understanding and improving local, complex, learning situations" [20].

Following works on ACAD, we understand that in order to achieve better learning outcomes, teachers need to carefully plan not only the content and forms of assessment but also take into consideration the learning environment. Therefore, we further consider "*activity*" as any engagement of students in the learning process (mental, physical, or emotional). The "*learning situation*" also comprises the three mentioned components above, as during hybrid seminars they were placed in different locations (onsite/online; in groups/individually; mixed locations).

The "*complex*" nature of the analyzed hybrid seminars, following the developed scenario based on the ACAD framework, is further interpreted from both experienced and future views. Within the scenario, the "complexity" of the studied learning situation was addressed through six main conceptual blocks: building understanding and connection, emotional support and resilience, enhancing empathy among peers, educational engagement, cultural awareness, and sensitivity. The offered scenario highlights the psychological, emotional, and communicative aspects of HFL, thereby providing a holistic learning environment.

The following tables represent the average number of hybrid sessions participants. Table I shows the number of participants at the first round of the seminar, while Table II reflects the number for the second round of seminars.

TABLE I. AVERAGE NUMBER OF HYBRID SESSIONS PARTICIPANTS (SEMINAR 1)

Participants (average number per session)	**Onsite**	**Online**
6 hybrid Sessions (2022/23)		
Students	13	17
Teachers	1	1
Guest Speakers	2	4

TABLE II. AVERAGE NUMBER OF HYBRID SESSIONS PARTICIPANTS (SEMINAR 2)

Participants (average number per session)	**Onsite**	**Online**
10 hybrid Sessions (Winter semester 2025)		
Students	10	8
Teachers	1	-
Guest Speakers	1	2

Following the ACAD framework, the paper represents a specific scenario for the hybrid classroom with the main focus on an unsafe region. The scenario includes epistemic, set, and social design. The epistemic design refers to the assignment that students received before the seminars and the activities they were supposed to be involved in during the hybrid session. At this stage, the topics for the seminars were carefully selected so that they would not raise negative emotions, but would encourage students to discuss sensitive topics and find solace or solutions in the suggestions provided.

The set design includes materials and platforms for interaction, presentation, sharing, and visualizing ideas, multimedia, etc.

The social design presupposes the planning of the ways students interact during the hybrid session. Ukrainian students from the frontline city have been living in conditions of social distancing for five years already. The feeling of alienation is exacerbated by disrupted relationships due to displacement, uncertain prospects for the future, and worries about their relatives' lives.

Hybrid seminars aimed not only to share knowledge on teaching English as a foreign language but also to establish networks and improve Ukrainian students' emotional state through peer collaboration in virtual settings. An opportunity to communicate with German students made Ukrainians feel that they were not outsiders struggling with their problems unsupported and that they belonged to a community that shared their values and had similar viewpoints.

Creating a learning design that considers all pedagogic properties can be facilitated through visualization. For such a purpose, the tool "Learning Designer" was used. Each step

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

of the hybrid sessions was specified in terms of learners' activities, teachers' involvement, duration, and resources to be used. The pie chart in Figure 1, which was generated from the information provided, illustrates the different types of learning and student interaction, allowing the teachers to analyze the effectiveness and patterns of students' participation and make adjustments before the seminars.

B. *Scenario of the Activity-Centered Analysis and Design (ACAD) Framework for Hybrid Classroom in an Unsafe Region*

1) *Background scanning*

Focus on Context: safety issues in the region were identified (e.g., due to electricity and Internet disruptions, the seminars could be rescheduled, and all necessary digital materials could be downloaded and were available offline).

Needs of Students and Teachers: social concerns and the emotional condition of the target audience, which can influence the class were considered (e.g., preparatory virtual phase for teachers – an advance meeting for planning the session; a constant channel for communication, such as Messengers, scanning students' psycho-emotional conditions, identifying potential hindering factors, and reporting these to a colleague teacher).

2) *Hybrid Classroom Design and Implementation Steps*

Flexible and adaptable: discuss possible adaptations in case safety conditions are violated (e.g., sharing video recordings and students' self-presentations on Flipgrid; recording voice messages and creating groups on Messengers for Session participants for instant communication).

Accessible and reliable: provide relevant technologies (open sources, free, and easy to use; e.g., create guest accounts for Miroboard, use Zotero as an open and free accessible digital library for sharing learning materials and enabling students to upload the materials themselves; an important option of offline access and asynchronous use of provided resources).

3) *Hybrid Classroom Learning Process Design*

Blended Learning: provide an option to learn both synchronously (live classes) and asynchronously (recorded lectures, online assignments);

Collaborating Learning Tools: interactive whiteboards, collaborating tools, such as Wooclap, Wordcloud, Mentimeter, etc., to facilitate students' interaction with an option of post-session access.

4) *Technologies Used*

Flowing Communication: despite interruptions and blackouts, students have access to supporting communication technologies (e.g., use of power banks to charge their gadgets), as well as accessible asynchronous learning materials.

Secured and certain: beware of cybersecurity issues and be ready to withstand the online threats (e.g., online support and instant messaging with volunteering IT specialists and/or IT competent students were at hand)

5) *Sustainable and Resilient Learning Environment*

Mental Well-being: offering workshops and activities, that enabled students to withstand their emotional strain (e.g., implementation of a slow-looking method, thinking routine methodologies from Project Zero developed by Harvard Graduate School of Education, integration of artful and pedagogical practices; addressing empathy and resilience with at-hand experiences; reflecting on students and teachers' own experiences).

Build Up the Community: interaction of students during collaborative projects, communication during and after the seminars, and extracurricular communicative activities (e.g., participation in the evening's Multilingual Speaking Club).

6) *Evaluation and Feedback*

Sustaining Improvements: regular meetings between teachers before and after the seminars, communication via emails and/or Messenger, helped participants feel supported and provided support to each other, enabling continuous modifications of the hybrid format seminars based on regular feedback from all parties involved (students, teachers, guest speakers).

Survey and Data Collection: short questionnaires during pre- and post-seminar phases with a flexible deadline were provided to all parties involved. The option for open questions enabled students to reflect on their own experiences, feelings, and concerns, as well as interests and ideas, thus shifting their role from the recipients of knowledge to the initiator and disseminator of self-authored seminar activities, learning materials, and scenarios.

IV. Main Findings and Discussion

The following section examines how hybrid classes can be organized effectively, based on the findings of our research.

A. *"Hybrid Harbors" as versatile collaborative spaces*

According to post-session survey data: "TEFL: Inter/transcultural learning and global education", the participants reflected on the commonality of their thematical foci, confirming creation within "hybrid harbors" of "a *collaborative learning environment* that encourages active participation, using hands-on activities and projects that promote authentic language use" (anonym. Session participant) and their ability to "demonstrate qualities like cultural awareness, empathy, and a willingness to engage in issues that transcend national borders" (anonym. Session participant).

The hybrid classroom enabled Ukrainian students to connect not only with peers but also with invited speakers from different countries. For example, an internally displaced Ukrainian speaker contributed to the session on *Multiperspective Representation of Cultures via Various Texts and Media*. The session on *Skills, Competencies, and Strategies in TEFL* with an intercultural focus featured a speaker from Spain. The seminar on *Arts and Pedagogy*, specifically the slow looking method in TEFL classes, was led by a speaker from the Czech Republic. A speaker from Canada contributed to the session on *Materials and Introducing data-driven EFL*, while a speaker from India

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

shared insights in the session on TEFL basics, focusing on transnational perspectives and interdisciplinary connections.

Such a versatile palette of speakers enabled students to collaborate beyond cultural and geographical borders. Their interactions were facilitated through synchronous discussions and interactions with the invited international guests using an interactive whiteboard, with continued access to the materials in the post-session phase. Thus, regardless of their physical location – whether internally displaced in Ukraine or based abroad – students could rely on consistently available learning resources and a safe, collaborative digital environment.

B. *Technological Challenges and Solutions*

Despite the limited availability of required technologies for efficient hybrid learning, students from Ukraine and their teacher devised the following practical solutions:

- Laptop "one for two" – students connected to the hybrid classroom using a shared laptop.
- Gadget as an additional asset – a personal cell phone or tablet was used for camera access or to join a collaborative online space.

At the same time, students physically present in Germany used their individual laptops in class. This setup enabled them to engage in synchronous written communication with their peers from Ukraine, teachers, and guest speakers while also interacting onsite with each other and their teacher, who was present in the classroom.

According to recent studies, the emotional intelligence of students, their psychological condition, and social performance are highly vulnerable. During the HFL sessions, the communicative gap widened, prompting teachers to introduce additional interactions, synchronous activities, and movement. In hybrid classrooms, where teachers must balance the demands of both onsite and online learners, German and Ukrainian educators highlighted that height-adjustable, movable tables represent a highly desirable condition. Such adaptable furniture supports effective multitasking, encourages dynamic interaction, and facilitates timely feedback across both teaching environments.

The target group of students faced distinct communicative challenges: Ukrainian students were coping with the realities of studying in an unsafe region, while German peers were experiencing hybrid learning of this nature for the first time – an eye-opening and unusual situation for many.

C. *Anthropocentric Challenges and Solutions*

To bridge the gap between live and online learners, teachers implemented the following communicative strategies:

- *Topic-to-go*- students were provided with relevant and innovative topics for the seminars, having received learning materials in advance and providing them in various formats (podcasts, texts for reading, short videos). Moreover, students were able to choose the topic of greatest interest before the session.
- *Collaborative tools*- platforms such as Mentimeter and Wooclap were used for synchronous polls, diagnostics of students' knowledge, and fostering group intreraction.
- *Digital dialogues*- group discussions, individual reflections, and flexible answer options allowed students to choose their preferred participation mode (e.g., chat contributions or whiteboard activities).
- *Q&A*- students were encouraged to pose questions, share short videos related to the seminar topic via screen sharing, and lead subsequent group discussions.
- *Synergized communities*- students identified shared perspectives during discussions, collaborated asynchronously on digital whiteboards, and co-created presentations or talks for upcoming seminars.

D. *Fostering SEL Through Creative and "Cozy" Hybrid Environments*

Following the educational model titled "Cozy corner, well-being, art, and human interaction" developed by the European University Alliance CHARM-EU, this approach highlights the importance of inclusive, emotionally supportive learning environments in hybrid classrooms [21]. It recognizes that physical space shapes learning experiences and that comfortable, inspiring settings – integrating elements such as plants, artwork, and soft furnishings – can nurture creativity, collaboration, and emotional well-being.

A central question arises: Can cozy corners exist in hybrid spaces? The answer lies in the very purpose of cozy corners – to create a supportive environment where students can overcome fear, discomfort, or other personal challenges and fully engage in learning. With thoughtful planning, this goal can indeed be achieved in hybrid settings as well.

The primary focus in hybrid classes should be on the learning tasks students are expected to engage in and on the ways students interact during these tasks to ensure a sense of emotional safety, comfort, and support. Based on the answers to questionnaires from the participants of seminar 2 (Appendix A. An Analytical Review of the Hybrid Learning Environment (Seminar 2): Assessing Benefits and Challenges; Table IV. Reflections on Hybrid Classrooms through the Lens of the ACAD Model), the background for the ACAD model implementation to create a "cozy environment" of the hybrid classroom is discussed (Table IV). These are some examples of learning activities that can be done in blended classes:

Journaling – students reflect on their mood and feelings at the beginning of the lesson, readiness to engage in learning activities, etc. At this stage, the teacher can introduce new vocabulary related to emotions that students can use and encourage them to share their ideas. However, it is worth remembering that if students have had some traumatic experience, the teacher should not force them to speak up – it is better to let them open gradually and engage in peer discussions only when they are ready, in order not to trigger any negative emotions. A digital journal might be a good solution, as it is easier to express yourself in a written

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

form without being judged or laughed at. If students feel at ease with their peers, they can record a short video using Flipgrid where they will describe their current mood and inspirations.

A *reading nook*, easily arranged in a physical classroom, may present some challenges in a hybrid setting. However, by setting aside time for quiet reading and allowing learners to choose their preferred reading space – whether at home or in the classroom – it is still possible to foster a similar sense of tranquility and reflective engagement. After reading, students can summarize the text in one or two sentences using a dedicated "write-it-out" space. Alternatively, the teacher can organize breakout rooms for students to discuss the story, share interpretations, and express their emotions. Carefully chosen texts often enhance the therapeutic and emotional impact of reading.

Learning new vocabulary through art. Art is widely recognized for its therapeutic value, particularly in helping individuals process trauma and express hidden talents. One effective way to engage students in vocabulary acquisition is through drawing. Teachers can encourage students to illustrate a word or phrase using digital tools such as Jamboard or Canva. This approach not only strengthens the connection between the visual image and the meaning of the word but also supports emotional well-being by allowing space for creative self-expression.

To develop the ability to see something positive even in minor events or activities, the teacher can create *a wall of gratitude* on Padlet, where students can share the acts of kindness they experienced. In addition to improving vocabulary and developing writing skills, it may also enhance faith in humanity and demonstrate that life is full of wonderful people.

A project-based task, "*My Comfort Box"*. This project-based task fosters connection, empathy, and a sense of value among students. The teacher invites learners to create a digital or physical box filled with various things: objects, images, even words they associate with comfort and safety. Students then briefly present their boxes in English, explaining the significance of each item. Afterwards, classmates are encouraged to respond by expressing their feelings about the presented items or by sharing similar personal experiences. This activity not only builds vocabulary related to emotions and personal well-being but also strengthens classroom cohesion and mutual understanding.

If carefully implemented, such activities can significantly improve students' self-perception, ability to analyze their emotional state, and help them find ways to address and overcome their fears, uncertainty, or other negative emotions. This is especially important for students who have to live and study in war conditions, who have been displaced, and who try to find themselves and their place in new surroundings. Through understanding, care, and empathy, EFL teachers can improve the lives of their students and make the world a better place to live.

E. Designing Emotionally Intelligent Hybrid Learning Spaces: A Three-Dimensional Framework

To implement "cozy corners" for both for online and onsite students, we adapted the ACAD model and the following design implementation steps (see Table III Educational Design Dimensions for Cozy Corners in Hybrid Classrooms based on ACAD Model). Offered "three-dimensional framework" follows the main structural and conceptual principles of the ACAD model, namely epistemic design, set design, and social design. To create meaningful hybrid learning experiences, there is a need for a thoughtful and well-balanced integration of physical and digital elements. This framework does not merely aim at increasing instructional effectiveness but also promotes psychological safety, emotional engagement, and a sense of belonging for all learners, regardless of their mode of participation. As a result, a holistic and emotionally aware learning environment is created.

Epistemic Design focuses on the tasks and activities students engage in. Onsite learners benefit from quiet collaboration, journaling, and peer coaching, while virtual participants reflect through digital journaling, breakout rooms, and asynchronous discussions. These activities are intentionally crafted to foster autonomy, emotional regulation, and metacognitive growth. Cozy corners, whether physical or virtual, serve as safe spaces for self-paced, reflective learning. These epistemic strategies are grounded in the understanding that learning is deeply intertwined with affect. When students are given structured opportunities to pause, reflect, and process, their cognitive engagement is enhanced. Moreover, the inclusion of emotional literacy tasks, such as identifying personal learning challenges or naming emotions, strengthens their ability to manage complex learning environments. The design thus moves beyond content delivery to include the cultivation of internal competencies necessary for lifelong learning.

Set Design addresses the tools and environments that support learning. Onsite settings include calming visuals, soft lighting, and sensory items, while virtual spaces offer mindfulness apps, music playlists, and digital lounges. It is essential to ensure a hybrid parity, the equal access to calming and reflective resources across modalities. The design should evoke comfort, inspiration, and engagement, helping students feel grounded and supported. The physical arrangement of the learning space – whether a quiet corner in a classroom or a customizable virtual interface – acts as a silent co-teacher. For onsite learners, cozy seating, nature imagery, and tactile tools like stress balls create a haven within an otherwise busy environment. For online participants, virtual backgrounds, personalized dashboards, or quick access to emotional support resources play a similar role. Importantly, hybrid learning platforms must be responsive to students' emotional and sensory needs, allowing for personalization and moments of quiet withdrawal when required.

Social Design emphasizes roles and interactions within the learning space. Facilitators encourage emotional check-ins and peer support, while students are empowered to self-

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

regulate and collaborate. Hybrid strategies, such as pairing onsite and online learners and using shared platforms like Padlet or Miro, promote inclusive and empathetic dialogue. Cozy corners become hubs for emotional connection and shared reflection, bridging the gap between digital and physical presence. This dimension particularly highlights the relational dynamics that shape emotional safety. Teachers act not only as knowledge facilitators but also as emotional anchors who model empathy and openness. Group norms, such as active listening, turn-taking, or the right to 'pass' during discussions, reinforce psychological safety. Through intentional structuring of social interactions, students learn to pass through interpersonal challenges, value emotional diversity, and build resilient learning communities.

Together, these elements create hybrid spaces that are not only functional but also emotionally intelligent and deeply human-centered. In times of crisis, such as during war or displacement, the value of emotionally responsive learning environments cannot be overstated. The integration of epistemic, set, and social design enables learners to engage not only cognitively but also affectively, fostering a sense of continuity, care, and connectedness that transcends physical boundaries.

F. ACAD and Tech pods

Social design elements, such as small group work, can increase students' sense of presence and belonging. These micro-communities within the classroom – whether physical or digital – serve as essential emotional anchors that foster both academic engagement and personal connection. For example, tech pods will be efficient in reducing the vulnerability of students' emotional intelligence, as well as their psychological condition and social performance. By offering structured yet flexible group settings, tech pods enable learners to express themselves more freely, engage in co-regulation of emotions, and practice interpersonal skills that are essential in emotionally intelligent learning spaces.

If to rephrase a well-known proverb, "The path to hybrid harbors passes through a techpod" (rephrased from *The path to Heaven passes through a teapot*). Following the CHARM Model of Hybrid Classroom [22], the implementation of *tech pods as group workstations* would facilitate students' active participation and enhance their self-performance, thus fostering emergent activities as an intersectional component of the ACAD model. This means that while set, epistemic, and social design elements operate independently, they converge within tech pods to create moments of synergistic learning. These emergent activities are not just by-products of design; they are indicators of a healthy, adaptive learning ecosystem.

To avoid feelings of frustration and loss, students will be able to engage in both physical and virtual collaboration. Such dual participation empowers learners to interact within multiple communication modes, thereby improving their digital fluency and social adaptability. Having arranged a controlled environment, another benefit is to reduce distractions and continue with focused group work [23]. These structured learning zones help students transition smoothly between cognitive tasks and emotional states, reinforcing self-regulation and collaborative resilience.

According to the hybrid session participants, a supportive and reliable environment is desirable for students to make them feel more secure and improve their emotional well-being. Feedback from students often highlights how emotionally safe tech pods feel – spaces where they are not only allowed but encouraged to pause, share, and co-create knowledge without fear of judgment. The consistency and continuity of developing students' *emotional intelligence* through facilitating their feelings and emotions via collaborative experiences established through tech pods is another advantage of this "socially constructing" technology. It reinforces emotional vocabulary, empathy, and trust, which are all crucial aspects of mental health literacy in academic contexts [7].

Social performance released during the collaborative activities also fostered teamwork spirit and communication skills. Having a balance between individual and group work, the students could "tailor" their individual learning scenarios. This balance supports differentiated instruction by allowing learners to negotiate their own pacing, depth, and mode of engagement, while still benefiting from the collective intelligence of the group.

Further implementation of the ACAD Framework to boost the efficiency of tech pods is offered as a consequential pathway to improving complex and challenging learning situations and designing learning-enhanced solutions. The framework acts as a blueprint for instructors, helping them to intentionally align the "design-time" elements (such as the choice of collaborative tools or pod configurations) with the "learn-time" dynamics (like interpersonal feedback and emotional co-regulation). In terms of the enhancement of tech pods workstations to develop students' emotional, psychological, and social needs, with a focus on both "design time" and "learn-time" directions, ACAD frameworks serve as a creative environment to promote learning and collaboration. This dual focus ensures that the learning space is not only effective at the moment of instruction but also adaptable to students' evolving needs and experiences over time.

Hybrid classes equipped with tech pods are uniquely designed to enhance students' emotional intelligence, psychological condition, and social performance. Following the principles of the ACAD framework, this study aims to provide a structured approach to designing these environments, ensuring they meet the diverse needs of students and promote holistic development. In this sense, tech pods are not merely logistical solutions for hybrid learning—they are intentional, emotionally intelligent design interventions that center the learner as a whole person.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

V. Conclusions and Perspectives

In this paper, we have outlined and pre-evaluated a range of both technological and anthropocentric design solutions, which will foster more effective implementation of hybrid classrooms as safe, immersive learning spaces. These approaches aim to improve both teachers' and students' physical well-being, psychological safety, and socio-emotional resilience. In particular, we emphasize the value of inclusive design, emotional intelligence, and collaborative learning environments.

While the current study is situated in the domain of teaching and learning English, the subject-specific aspects are not the only focus. A central contribution lies in considering the situational and social dimensions of hybrid learning, which are transferable to other disciplines. The Activity-Centered Analysis and Design (ACAD) framework emphasizes physical, epistemic, and social aspects of learning, making it adaptable beyond language education. Thus, although the examples stem from EFL contexts, the methodological principles and attention to social-emotional competencies can be generalized to other subject areas. The following structural and pedagogical elements have emerged as key recommendations for future implementation:

- Classroom infrastructure that refers to the Toolkit on Accessibility from UNICEF, including spacious layouts, barrier-free access for students and teachers with disabilities (e.g., wheelchair users), appropriate lighting, low-noise soundscapes, and temperature control to ensure inclusive, health-supportive environments.
- Height-adjustable, mobile teaching stations such as pneumatic rolling desks that allow instructors to adapt their physical positioning in the space, thereby reducing physical strain and encouraging fluid classroom management.
- Node-style seating and mobile tablet armchairs, which enable students to form learning clusters quickly and reconfigure the physical space for collaborative group work. Such flexible furniture promotes autonomy and peer interaction in onsite settings.
- "Cozy corners" (online and onsite) as hubs of emotional connection, mindfulness, and shared reflection. These intentionally designed spaces allow for self-regulation, metacognitive pauses, and moments of comfort in otherwise demanding hybrid environments.
- Tech pods for students, as outlined in the CHARM-EU Hybrid Classroom Model, equipped with multiple microphones, cameras, and screens to provide seamless access to group discussions and collaborative work, ensuring equal inclusion for both online and onsite students.
- Targeted training for teachers in the HFL classroom technology implementation and digital-emotional classroom management strategies. This includes designing open, friendly, and inclusive learning environments, as well as adopting flexible instructional models that accommodate the needs of both remote and in-person learners.

These elements collectively work toward a human-centered vision of hybrid education that places equal value on emotional well-being and cognitive achievement.

Looking ahead, further research is necessary to deepen our understanding of how socio-emotional interactions can be optimized in HFL environments. Teacher professional development will remain central – particularly in the fields of socio-emotional learning, empathy-building, intercultural awareness, and trauma-informed pedagogy. Existing studies emphasize that educators require specific training to support learners from unsafe regions, post-conflict zones, and disrupted educational trajectories [18, 19, 24, 25, 26-28, 29-30].

As the demand for hybrid education models continues to grow, especially in contexts disrupted by instability, displacement, or infrastructural limitations, it becomes increasingly urgent to expand pedagogical frameworks that are emotionally intelligent, inclusive, and dignity-affirming. Such approaches are essential to cultivating resilient learning communities capable of responding to global challenges with agency and care.

Currently, the research is expanding to include more transnational participants, sharing knowledge globally. In 2025, the course "Transcultural Project-Based Learning. Multilingualism through the Arts" will be offered continuously to German and Ukrainian students in a hybrid format. The course will focus on educational project management, intercultural communication, and interdisciplinary collaboration. Students will also participate in the international eTwinning project BLABL.ART, partnering with institutions from Italy, France, Reunion, and the Czech Republic. The course will be guided by the core principles of universality, interdisciplinarity, openness, flexibility, respect, and resilience, ensuring continuity with the human-centered approach to hybrid education at Julius-Maximilians-Universität Würzburg and other partners.

The proposed concept of "hybrid harbors" – understood as emotionally intelligent, safe, and human-centered hybrid learning spaces – holds the potential to be adapted across a wide range of challenging learning contexts. These may include refugee and displacement education programs, remote and rural communities lacking digital infrastructure, post-pandemic recovery classrooms, and regions affected by ecological crises. The anthropocentric emphasis of such models positions learners and educators as active agents of empathy and transformation, rather than passive recipients of content. Building frameworks of inclusivity, emotional safety, and transdisciplinary collaboration is essential in cultivating a new generation of global citizens who can

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

respond with resilience and creativity to complex societal disruptions.

Ultimately, the pedagogical designs and concepts introduced here do not only aim to ensure educational continuity. They aspire to create shared spaces of healing, connection, and empowerment through intercultural dialogue and artistic expression. In so doing, they open up possibilities for a new kind of education—one that is not only innovative and resilient but also profoundly human.

ACKNOWLEDGMENT

We are deeply grateful to Dr. Matthias Erhardt, Managing Director of the Professional School of Education at Julius-Maximilians-Universität Würzburg, for his continuous support and valuable advice, which have significantly contributed to the success of the initiative.

REFERENCES

[1] N. Lazebna and K. Lut. "Hybrid Harbors: Immersive Learning Spaces for Unsafe Regions". eLmL 2025: The Seventeenth International Conference on Mobile, Hybrid, and On-line Learning. Published in ThinkMind Digital Library. ISBN: 978-1-68558-271-5. pp. 1-6, 2025

[2] J. Britton. "Virtual, remote and hybrid checklist". TEDx Derry Londonderry Studio, Aug. 2021, YouTube. https://www.ted.com/talks/jennifer_britton_virtual_remote_and_hybrid_checklist_jan_2021?language=en

[3] Competence Centre for Digital Teaching and Learning, Universität Würzburg. "DigiPhilLab". In Universität Würzburg, n.d. Universität Würzburg. https://www.uni-wuerzburg.de/pse/ueber-uns/kompetenzzentrum-fuer-digitales-lehren-und-lernen/raeume-fuer-digitales-lehren-und-lernen/digiphil/

[4] N. Lo. "Emotional bridge in higher education: Enhancing self-efficacy and achievement through hybrid engagement". In *ESP Rev*, 5, pp. 7–23, 2023.

[5] M. R. Griffard, M.E. Marraccini, D. Barber, L. Sartain & D.C. Griffin. "'This Is Not Online Education. It's Disaster Education': A Qualitative Study of Remote and Hybrid School Climates During COVID-19". In *AERA Open*, 11, Article 23328584251338058, 2025. https://doi.org/10.1177/23328584251338058 (Accessed: 21 November 2025)

[6] E. Pipia and L.R. Williams. "Navigating the "New Normal" in Higher Education: Interplay between Social Disruptions, Academic Achievements, and Emotional Well-Being". In *INTED2024 Proceedings*, pp. 6649–6656, 2024. IATED. DOI:10.21125/inted.2024.1742 (Accessed: 21 November 2025)

[7] L. Conwi, F. Pinar, and J. Destura. "Exploring mental wellbeing, distress and adjustment in a blended learning environment". In *Journal of Interdisciplinary Perspectives*, 2(7), pp. 146–157, 2024. https://doi.org/10.69569/jip.2024.0148 (Accessed: 21 November 2025)

[8] M. Eisenmann, C. Ludwig, T. Summer, D. Becker, and N. Krüger. "Mental Health in English Language Education - An Introduction". In Eisenmann, M., Ludwig, C., Summer, T., Becker, D. & Krüger, N. (Eds.). *Mental Health in English Language Education*, pp. 9-30. Narr., 2024.

[9] J. Petry, K. Neubert, J. Duda, J. Krämer, J. Trippler, C. Oeltjebruns, and A. Engel. "Hybrid activity-orientated learning spaces to strengthen social-emotional learning processes in the social, therapeutic and health sciences". In *INTED2025 Proceedings*, pp. 4292–4301, 2025. IATED. doi: 10.21125/inted.2025.1085 (Accessed: 21 November 2025)

[10] R.L. Geesa, K. Robbins, and K. Shively. "The collaborative model for teaching o-SEL: Preparing educators to design online environments for social-emotional learning". In *Journal of Online Learning Research*, 8(1), pp. 67–100, 2022.https://www.learntechlib.org/primary/p/220643/. (Accessed: 21 November 2025

[11] S.L.Moore. *SEL at a Distance: Supporting Students Online* (Social and Emotional Learning Solutions), 2021. WW Norton & Company. ISBN 978-1-324-01658-8

[12] P. Sharma and J. Bhasin. "Well-being of faculty members in era of hybrid education: A way forward". In *Enhancing Productivity in Hybrid Mode: The Beginning of a New Era*, p. 372, 2023. https://www.researchgate.net/profile/Atul-Kumar-135/publication/371984403_Impact_of_Web_Technology_Marketing_on_Consumer_Durable_Items/links/64a12ed2b9ed6874a5ed3097/Impact-of-Web-Technology-Marketing-on-Consumer-Durable-Items.pdf#page=387 (Accessed: 21 November 2025

[13] T.C. Philibert and A. Slade. *Everyday SEL in the Virtual Classroom : Integrating Social Emotional Learning and Mindfulness into Your Remote and Hybrid Settings.* New York, New York; Routledge, 2022. Print. doi:10.4324/9781003141068 (Accessed: 21 November 2025

[14] N.J. Mentzer, T.M. Isabell, and L. Mohandas. The impact of interactive synchronous HyFlex model on student academic performance in a large active learning introductory college design course. *J Comput High Educ* 36, pp. 619–646, 2024. https://doi.org/10.1007/s12528-023-09369-y (Accessed: 21 November 2025

[15] R. S. Levine, R.J. Lim, and A.V. Bintliff. "Social and emotional learning during pandemic-related remote and hybrid instruction: Teacher strategies in response to trauma". In *Education Sciences*, 13(4), Article 411, 2023. https://doi.org/10.3390/educsci13040411 (Accessed: 21 November 2025

[16] E.A. Meland and G. Brion-Meisels. "An integrative model for culturally sustaining SEL in the classroom". In *Social and Emotional Learning: Research, Practice, and Policy*, Article 100042, 2024. https://www.sciencedirect.com/science/article/pii/S2773233924000160?via%3Dihub (Accessed: 21 November 2025

[17] F. Karaferye. "Investigating teachers' experiences with digital classroom management and incorporating social and emotional learning". In *Turkish Online Journal of Distance Education*, 25(1), pp. 179–199, 2024. DOI:10.17718/tojde.1210701 (Accessed: 21 November 2025

[18] M. S. Alkhowailed et al. "Digitalization plan in Medical Education during COVID-19 Lockdown". Informatics in Medicine Unlocked, vol. 20, pp. 1–6, 2020. https://doi.org/10.1016/j.imu.2020.100432. (Accessed: 21 November 2025

[19] E. Kalmar, T. Aarts, E. Bosman, E. Beinema, R. Winkel, and M. van der Sanden. "The COVID-19 Paradox of Online Collaborative Education: When you cannot physically meet, you need more social interactions". Heliyon, vol. 8 (1), pp. 1–15, 2022. https://doi.org/10.1016/j.heliyon.2022.e08823 (Accessed: 21 November 2025

[20] P. Goodyear, L. Carvalho, and P. Yeoman. "Activity-Centred Analysis and Design (ACAD): core purposes, distinctive qualities and current developments". Educational Technology Research and Development, vol. 69(2), pp. 445-464, 2021. https://doi.org/10.1007/s11423-020-09926-7 (Accessed: 21 November 2025

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[21] A. Gilmore, T. Daher, and M. Peteranetz. "Multi-year case study in blended design: student experiences in a blended, synchronous, distance controls course". Computers in Education, vol. 12(1), 2021. https://coed.asee.org/2021/03/31/multi-year-case-study-in-blended-design-student-experiences-in-a-blended-synchronous-distance-controls-course/. (Accessed: 21 November 2025

[22] CHARM-EU. Resource center, 2022. https://charm-eu.eu/resources/resource-center/

[23] L. Angelone, Z. Warner, and J.M. Zydney. "Optimizing the technological design of a blended synchronous learning environment". Online Learning, vol. 24(3), pp. 222–240, 2020. https://olj.onlinelearningconsortium.org/index.php/olj/article/view/2180/987 (Accessed: 21 November 2025

[24] S. Lakhal, J. Mukamurera, M.-E. Bédard, G. Heilporn, and M. Chauret. "Features fostering academic and social integration in blended synchronous courses in graduate programs". International Journal of Educational Technology in Higher Education, vol. 17(1), pp. 1–22, 2020. https://doi.org/10.1186/s41239-020-0180-z. (Accessed: 21 November 2025

[25] A. Raes, L. Detienne, I. Windey, and F. Depaepe. "A systematic literature review on synchronous hybrid learning: gaps identified". Learning Environments Research, vol. 23, pp. 269–290, 2020. https://doi.org/10.1007/s10984-019-09303-z. (Accessed: 21 November 2025

[26] S. Annamalai, A. Vasunandan, and A. Mehta. "Social isolation and loneliness among Generation Z employees: can emotional intelligence help mitigate?" Cogent Business & Management, vol. 12(1), 2024. https://doi.org/10.1080/23311975.2024.2441474. (Accessed: 21 November 2025

[27] UNICEF. Accessibility Toolkit, 2022. https://accessibilitytoolkit.unicef.org/

[28] L. de Wal Pastoor. "Reconceptualising refugee education: exploring the diverse learning contexts of unaccompanied young refugees upon resettlement". In Refugees, Interculturalism and Education, pp. 37-58, 2020. Routledge. https://www.tandfonline.com/doi/abs/10.1080/14675986.2017.1295572 (Accessed: 21 November 2025

[29] C. Koehler, N. Palaiologou, and O. Brussino. "Holistic refugee and newcomer education in Europe: Mapping, upscaling and institutionalising promising practices from Germany, Greece and the Netherlands", OECD Education Working Paper No. 264, 2022. https://one.oecd.org/document/EDU/WKP(2022)2/en/pdf (Accessed: 21 November 2025

[30] N. Lazebna, K. Lut, and E. Dieser. "Inclusion as Solution: Integration Challenges of Ukrainian Refugee Students in the Educational System of Germany". In Education Leadership in the Shadow of Wars, ed. By M. Gutman, pp. 113-122, Routledge, Taylor and Francis, USA, 2024. https://www.taylorfrancis.com/books/edit/10.4324/9781003571575/challenges-facing-education-leadership-shadow-war-mary-gutman. (Accessed: 21 November 2025

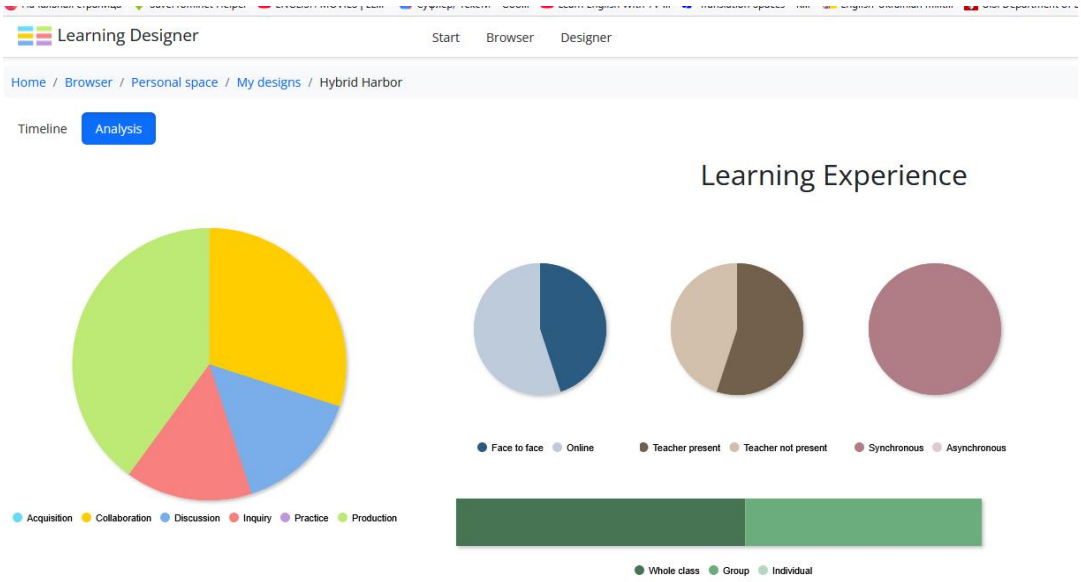


Figure 1. Visual representation of the learning design for hybrid sessions

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Table III. Educational Design Dimensions for Cozy Corners in Hybrid Classrooms based on ACAD Model

Dimension	Focus	Application	Design Implications
1. Epistemic Design (Tasks and Learning Activities)	Learning tasks and activities students engage in	Onsite: Reading, journaling, peer coaching, quiet collaboration Virtual: Digital journaling, breakout rooms for reflection, asynchronous discussion boards	Tasks should foster autonomy, emotional regulation, and metacognitive skills. Cozy corners support self-paced, reflective, and emotionally aware learning.
2. Set Design (Physical and Digital Environment)	Tools, spaces, and resources available	Onsite: Comfortable seating, soft lighting, calming visuals, noise-reducing elements; tools like notebooks, tablets, sensory items Virtual: Calming breakout rooms, virtual lounges; tools like mindfulness apps, music playlists, journaling platforms	Ensure hybrid parity by offering equivalent calming and reflective resources in both physical and digital formats. Design should evoke comfort, inspiration, and engagement.
3. Social Design (Roles and Interactions)	Social arrangements and participant roles	Facilitator: Promote use of cozy corners for emotional check-ins and peer support Students: Self-regulate, support peers, reflect on learning Hybrid Strategy: Pair onsite and online students for shared reflection; use collaborative tools like Padlet or Miro	Encourage inclusive, emotionally intelligent interactions. Cozy corners become spaces for empathy, collaboration, and shared reflection across modalities.

Table IV. Reflections on Hybrid Classrooms (Seminar 2) through the Lens of the ACAD Model
(*The responses provided by participants were retained in their original form, including any grammatical errors, stylistic inconsistencies, or language issues, to preserve the authenticity and integrity of their input)*

ACAD Dimension	Focus	Participant Reflections	Design Implications
Epistemic Design (Tasks and Learning Activities)	Learning tasks and emotional engagement	"The multiple informative resources our instructor provided us with. I really kept me engaged and curious." "I particularly liked that we got to know lots of different activities and links to sites with even more information on classroom activities and I'm eager to try them out in the classroom -> artful thinking project for example".	Design tasks that promote emotional expression, autonomy, and metacognitive awareness. Include reflective journaling and structured check-ins.
Set Design (Physical and Digital Environment)	Tools, spaces, and sensory elements	"Creating a collaborative art project, such as a jounal with labels in multiple languages, promotes practical vocabulary learning and cultural expression in a creative way." "Integrating transmedial creative	Create calming, inclusive environments both onsite and online. Use sensory tools, soft lighting, and digital mindfulness resources (e.g., Thinking Routine Toolbox by Project Zero, Harvard

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

ACAD Dimension	Focus	Participant Reflections	Design Implications
		environments into TEFL by combining storytelling, social media simulations, AR/VR, and multimedia projects where students create and adapt content across formats (e.g., blogs, videos, comics) to build language skills and digital literacy.”	Graduate School of Education).
Social Design (Roles and Interactions)	Relationships, collaboration, and support	“I got a new friend from Germany due to the classes. So it was definitely incredible experience especially in practicing foreign language”. “I've also learned how to create a collaborative learning environment that encourages active participation, using hands-on activities and projects that promote authentic language use.”	Foster emotionally intelligent interactions. Use hybrid pairings and collaborative platforms like Padlet or Miro to bridge modalities.

Appendix A

An Analytical Review of the Hybrid Learning Environment (Seminar 2): Assessing Benefits and Challenges
(*The responses provided by participants were retained in their original form, including any grammatical errors, stylistic inconsistencies, or language issues, to preserve the authenticity and integrity of their input)*

Section I: Overall Experience. Evaluating the Balance Between Accessibility and Reduced Engagement in Hybrid Classrooms

Positive Reflections:

- "No problems or challenges, everything was well-organised!" (Anonymous 1)
- "I wouldn’t redesign anything." (Anonymous 7, Redesign section)
- "Everything was wonderful!" (Anonymous 1, Support section)
- One participant described the experience as "interesting and exciting" (Anonymous 1, Emotional Well-being section) and another gained "a new friend from Germany" (Anonymous 3, Emotional Well-being section).

Challenges and Negative Reflections:

- "at first it was a bit awkward because you feel people watching you on the big screen, but then you get used to it. The biggest challenge was that at times you feel separate from the others, because when there is a discussion offline, you can't always hear what is being said, so you don't have the confidence to add something of your own." (Anonymous 2)
- "Tech issues" (Anonymous 7)
- "even though the others were present vie zoom, the lesson felt kind of empty with only 5 on-sight participants." (Anonymous 8)
- "Disconnected when during the zoom class my headphones didn’t work" (Anonymous 6, Engagement section)
- "you generally felt disconnected from the class due to the fact that the majority of the participants was attending online." (Anonymous 8, Engagement section)
- "I prefer an on-site class format. Hybrid always creates more distance." (Anonymous 5, Belonging section)

Section II. Key Benefits of the Hybrid Learning Model: Prioritizing Flexibility and Accessibility

Flexibility and Convenience:

- "Flexibility, while I was able to join remotely" (Anonymous 2)
- "the possibility to participate from home in case of unexpected events / sickness / etc. ..." (Anonymous 4)
- "lowered it [stress], knowing I could join online" (Anonymous 9, Emotional Well-being section)

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

- "The hybrid format had a positive impact on my emotional well-being because it gave me flexibility. I could learn from home, which reduced stress related to commuting and saved time." (Anonymous 7, Emotional Well-being section)
- "This format is convenient for presentations, videos, etc., and participants both in the classroom and online can easily be in context." (Anonymous 3)

Global Participation:

- "I guess, such format is really great, bc students from around the world can participate in this course" (Anonymous 3)

Accessibility for Specific Needs:

- "The possibility to see the presentations looking at my laptop's screen. I can see all the words and pictures clearly despite my poor eyesight" (Anonymous 7)

III. Challenges and Areas for Improvement: Bridging the Online-Offline Divide

Communication Gaps & Disconnection:

- "The biggest challenge was that at times you feel separate from the others, because when there is a discussion offline, you can't always hear what is being said, so you don't have the confidence to add something of your own." (Anonymous 2)
- "I sometimes felt disconnected due to poor sound quality (because offline students sat far from speakers)." (Anonymous 7, Engagement section)
- "Not being able to really communicate with the online group during the class." (Anonymous 4, Isolation section)

Sense of Belonging/Inclusion:

- "The moment when I felt connected was when I could share my own teaching experience." (Anonymous 3, Belonging section)
- "There was no sense of belonging as we were only 5 people attending in present." (Anonymous 8, Belonging section)
- "Overall, I felt included in the class, at least I felt like I belonged to the offline group. But I don't know how the online group felt." (Anonymous 15, Belonging section)
- One participant, however, felt the format *helped* inclusion: "The hybrid format actually helped me feel included because it gave everyone a chance to participate, no matter where they were." (Anonymous 7, Belonging section)

On-site Attendance Concerns:

- "seeing as the on-sight attendance for the german students was mandatory, there really was nothing to "work well" for us." (Anonymous 13)
- "i would like to see a minimum of on-sight participants reqiurement for the course." (Anonymous 8, Redesign section)

IV. Collaboration and Engagement Strategies

Collaboration Tools:

- "Chatting, Zoom meeting, whatsapp group" (Anonymous 9, Collaboration section)
- "We worked with google docs and communicated vie Whatsapp. I'd say the communication was the easy part." (Anonymous 8, Collaboration section)
- "Beyond classes, we collaborated using WhatsApp and Zoom. It worked for all of us since it's convenient way to reach out to each other and discuss the" (Anonymous 17, Collaboration section)
- "We managed group work quite effectively by using online tools for communication and collaboration, such as shared documents, Zoom sessions and chat groups." (Anonymous 7, Group Work section)

Engagement Strategies:

- "interest in the topic, seeing the goal" (Anonymous 9, Engagement section)
- "The Zotero platform was very convenient because all materials and homework were easy to access. The course included a lot of practical tasks, which kept me motivated." (Anonymous 7, Engagement section)
- "Use of presentations and videos." (Anonymous 2, Engagement section)
- "The multiple informative resources our instructor provided us with. I really kept me engaged and curious." (Anonymous 4, Engagement section)
- "Presentation of Methods and Tools, practical applications inside the class to use in or for the project, motivated me and helped me stay engaged" (Anonymous 5, Engagement section)
- "Thinking about the final project and picking up ideas during the seminar" (Anonymous 6, Engagement section)

V. Instructor's Role: Primarily Supportive

Key Support Mechanisms:

- "Our teacher was always ready to answer out questions" (Anonymous 3)
- "The instructor was so nice. She wrote us emails, gave us inspiration about project, and asked all our questions about that" (Anonymous 1, Instructor Support section)

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

- "The instructor always took into account the involvement of all participants, repeated questions if necessary, or explained twice when the idea was not understood." (Anonymous 12, Instructor Support section)
- "The instructor was very supportive throughout the course. She gave clear instructions on how to make the final project. She encouraged participation, asked for feedback and provided helpful examples connected to art." (Anonymous 7, Instructor Support section)
- "Our instructor supported our learning experience by creating our own zotero library where she shared our course's content." (Anonymous 18, Instructor Support section)
- "When posing a question to the class, she also tried to include the online group in the discussion, which helped a bit with the collaboration between the two groups (online & offline) I think." (Anonymous 19, Instructor Support section)

Minor Suggestions for Improvement:

- "Maybe a little bit more encouragement / guidance from the instructor and more exact instructions for the final project." (Anonymous 4, Support section)

VI. Emotional Well-being and Stress Levels

The impact on emotional well-being varied, but for many, the flexibility offered by the hybrid format was a stress reducer. Initial nervousness was often overcome by increased familiarity and confidence.

Stress Reduction through Flexibility:

- "lowered it, knowing i could join online" (Anonymous 9, Emotional Well-being section)
- "It generally helped because I was more flexible and did not need to meet at a fixed spot to collaborate with the others" (Anonymous 5, Emotional Well-being section)
- "The hybrid format had a positive impact on my emotional well-being because it gave me flexibility. I could learn from home, which reduced stress related to commuting and saved time." (Anonymous 7, Emotional Well-being section)

Initial Nervousness and Developing Confidence:

- "This format brought a bit of stress at the beginning, but it gave me experience, and now it doesn't seem scary." (Anonymous 2, Emotional Well-being section)
- "At the beginning of the course, I felt both excited and slightly nervous because it was a new format for me. By the end of the course, I felt more confident, inspired, and motivated." (Anonymous 11, Emotional State section)

Occasional Stressors:

- "I was slightly more stressed just before the draft presentation, because I wasn't sure if I understood the project assignment correctly." (Anonymous 4, Emotional State section)
- "at the end i was a bit more stressed because of the final presentation coming closer" (Anonymous 6, Emotional State section)

VII. Suggestions for Future Hybrid Classes

Improving Communication and Interaction:

- "Maybe settle from the beginning of the course on a way of communication between the online and offline participants, so that we can exchange ideas more easily and establish a connection from the start." (Anonymous 12, Suggestions section)
- "I guess that the group with all participants on networks is the best way to improve communication" (Anonymous 20, Suggestions section)

Technical and Structural Improvements:

- "It would be good to try to display all the information, including questions, on the screen, as this prevents online participants from not being involved." (Anonymous 10, Redesign section)
- "Make the order and planning more rigid. Provide more exact guidelines for what to do in order not to be confusing." (Anonymous 9, Suggestions section)
- "I would like to see a minimum of on-sight participants reqiurement for the course." (Anonymous 8, Redesign section)

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Consequences of EU Data Regulation for Smart Home Data Economy: Extending Analysis with a Case Study

Paul Seidel, Felix Fischer, Dirk Labudde
Faculty Applied Computer Sciences and Biosciences
Hochschule Mittweida University of Applied Sciences
Mittweida, Germany
e-mail: {seidel6 | fische11 | labudde}@hs-mittweida.de

***Abstract*—The entry into force of the European Union (EU) Data Act 2024 creates new opportunities for the European data market, but also new challenges. One such challenge is the parallel application of the EU General Data Protection Regulation (GDPR). It is therefore necessary to analyse these two regulations and their consequences for the stakeholders in the Smart Home sector. To this end, we have researched the Smart Home sector and its stakeholders and identified potential conflicts between the EU Data Act and the EU GDPR. One such conflict arises in the management of personal data from multi-user environments. In the Smart Home in particular, several users share different devices, such as smart TVs and thus generate mixed data sets that are not compliant with the regulation. If a member of the user community wishes to transfer their data to a third party in accordance with their rights guaranteed by the EU Data Act, the third party must be able to ensure that the transferred data is not also the personal data of another user. We supplement our regulatory and technical analysis with a case study of the Bosch Smart Home Controller, which demonstrates practical approaches and unresolved issues in managing personal and non-personal data, enforcing GDPR provisions, and realizing the potential of the Data Act in actual home automation deployments.**

Keywords-EU data act; gdpr; contradiction; smart home.

I. Introduction

This extended paper builds on our earlier study of the consequences of the EU Data Act and GDPR for the modern smart home data economy, originally presented at IARIA ICDS 2025 [1]. This version supplements the regulatory analysis with a focused case study examining real-world data security and privacy practices implemented in a modern Smart Home Controller API.

Increasing digitalisation and the steady expansion of data-based business models have placed the so-called data economy at the heart of economic and technological developments. Data is regarded as the new oil of the 21st century [2] and is essential for value creation in areas such as machine learning, whose economic potential through generative models has recently been estimated at several trillion dollars [3, p. 3]. This makes the regulation and utilisation of data a key challenge for modern economies.

The Smart Home plays a central role in this data economy, as connected devices generate a wide range of data that can benefit both users through better services and manufacturers through commercial exploitation [4]. Currently, however, European consumers often do not have access to their data, which hampers innovation and competition [5]. The European Union (EU) Data Act aims to redress this imbalance by granting users extensive rights to their data, including real-time access to device-generated data [6, Art. 3]. In addition to the General Data Protection Regulation (GDPR), which already provides for the right to data portability [7, Art. 12 para. 3], the Data Act aims to promote competition and interoperability, for example through cloud switching and access to manufacturer data for repair services [6, Recital 78]. At the same time, however, there is a tension between the two regulations. While the Data Act also requires the transfer of mixed data sets, the GDPR prioritises the protection of personal data and prescribes strict sanctions for violations [6, Art. 1 para. 5]. This leads to legal uncertainties, particularly in the Smart Home, where mixed data sets are often created. With the EU Data Act becoming applicable law in September 2025, this issue is becoming increasingly relevant and requires technical solutions to take into account both regulatory requirements and the technical innovation potential.

The urgency of this study arises from that recent entry into force of the EU Data Act, which significantly reshapes the regulatory landscape for data access and sharing in Europe. Particularly in Smart Homes, where multiple users often interact with interconnected devices and generate mixed datasets, the practical application of the Data Act introduces tensions. This study examines these tensions, focusing on the legal and technical challenges of managing personal data in multi-user environments and ensuring regulatory compliance. The specific designs and implementations of the technical and legal solutions to these challenges are beyond the scope of this study.

A recent U.S. decision in the antitrust case against Google further highlights a principle that closely resonates with the EU Data Act: access to data is essential for fair competition. By obliging Google to share parts of its search data with competitors, the court recognized that control over large datasets can entrench market dominance and stifle innovation [8]. The EU Data Act builds on the same insight, seeking to open up data flows in sectors such as the Smart Home by ensuring that users can access and share the information generated by their devices. Both measures rest on the idea that unlocking data for wider use creates opportunities for new entrants and service providers, thereby fostering more dynamic and competitive digital markets. This frames the facilitation of fair access to

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

data as a global and international issue.

After this introduction, the key stakeholders and challenges in the Smart Home sector are discussed in Section II, focusing on their interests and the inherent problems in this environment. Also in Section II, the concept of the Smart Home is defined, and the roles of relevant stakeholders are explored. Section III then examines the challenges posed by the data economy in Smart Homes, particularly data protection issues and power asymmetries between consumers and service providers. In Section IV, the EU Data Act and the GDPR are analyzed, providing an overview of both regulations, highlighting potential conflicts, and assessing their implications for the Smart Home sector. Section V discusses enforcement trends of the GDPR and their broader implications for compliance, while Section VI presents a case study of Bosch's Smart Home Controller as a practical example of regulatory challenges in implementation. Finally, the paper concludes with a summary of findings and directions for future work in Section VII.

II. Stakeholders and Problems in the Smart Home

The Smart Home sector is a central component of the modern data economy, in which consumers, device manufacturers and service providers operate in a complex network of economic and regulatory relationships. In order to better understand the opportunities and risks of this sector, a comprehensive analysis of the players involved and their interests is required. Therefore, the basic concepts and functioning of the Smart Home, as well as the roles of the relevant players are first examined and the potentials and risks arising from the use of the data generated in the Smart Home are analysed. The central challenges of the data economy in the Smart Home context are then analysed, with a focus on data protection problems, economic effects and power imbalances between users and providers.

A. Smart Home - Definition and concept

The Smart Home is based on the technologies of Embedded Systems and the Internet of Things (IoT). Embedded Systems are specialised, integrated computer systems designed to automate and simplify the functionalities of the host device [9, ch. 1]. Networking via technologies such as IEEE 802.11, usually referred to as WLAN, or Bluetooth creates the IoT, which enables communication between devices and their real-time interaction [10].

In the Smart Home, this IoT architecture is used to network household appliances and automate everyday processes [11]. The aim is to increase comfort, safety and efficiency, for example through smart thermostats that optimise energy consumption based on the habits of the residents [12][13].

An overview of the terms Embedded Systems, Internet of Things and Smart Home is summarised in Table I.

TABLE I
Definition of terms: Embedded Systems; IoT; Smart Home

Term	Definition	Examples
Embedded systems	mechanical and electrical systems with integrated software	modern cars, cash register systems, ATMs
Internet of Things (IoT)	interconnected embedded systems	Industry 4.0, car-to-car communication
Smart Home	IoT systems in home automation	vacuum/mopping robots, SmartTVs

The networking of Smart Home devices opens up numerous opportunities to make everyday life easier and to organise processes more efficiently by networking various household appliances. For example, the data collected can be used to increase comfort and energy efficiency in households [14]. In the healthcare sector in particular, wearables, such as smart watches offer the possibility of recognising medical emergencies, such as strokes or heart attacks at an early stage, enabling faster and potentially life-saving interventions [15][16]. Home automation also benefits the environment, as intelligent control systems can optimise the energy consumption of appliances. Automated adjustment of heating, lighting and other systems not only reduces costs for residents, but also helps to reduce the ecological footprint [17][18]. An illustrative example is the automatic switch-off of radiators as soon as windows are opened [19].

Smart home data is also used to improve building security. Intelligent monitoring systems can recognise break-ins at an early stage and initiate preventative measures. In addition to traditional dangers, such as burglaries, invisible risks, such as voltage peaks, critical air conditions or structural damage to buildings can also be detected and communicated to the residents [20, p. 7][21][22]. In addition, personalised functions, such as alarm clocks, music or news, enrich daily life by being tailored to the individual preferences of residents. The use of modern AI technologies makes it possible to analyse the collected data and turn it into meaningful automated decisions [23][24].

The symbiosis between AI and Smart Home technologies mutually reinforces both areas. AI relies on using large amounts of data to develop powerful models, while Smart Home devices continuously generate such data [25][26]. This creates a market in which data trading plays a central role. Companies that rely on AI-supported solutions buy the necessary data, while the owners of this data can monetise it. This creates an economic incentive, especially for companies without the technical resources to utilise their own data [6, Recital 19].

Networked systems also offer potential at a societal level, for example in public health management. During the COVID-19 pandemic, it became clear how valuable data-driven systems can be in tracing chains of contact [27]. Applications such as the *Corona-Warn-App* [28] helped to break chains of infection and slow down the spread of the virus. At the same time, however, the collection and processing of sensitive data raises questions about the protection of privacy. While the pandemic has demonstrated the benefits of such technologies, it has also revealed the risks of large-scale data collection

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

and storage. Sensitive information on health, behaviour and habits could be exposed or misused by cyberattacks, causing considerable harm to the individuals concerned [29, pp. 11]. Weighing up the benefits of data-driven innovations against the risks to privacy and security therefore remains a key challenge for the Smart Home data economy.

B. Relevant stakeholders and interests

To analyze the dynamics within the Smart Home ecosystem, stakeholders were grouped along two dimensions: their level of access to data and their technical know-how to generate value from it. Four central stakeholder groups (SG) operate in the Smart Home sector. The relationships between the stakeholder groups are shown graphically in Figure 1.

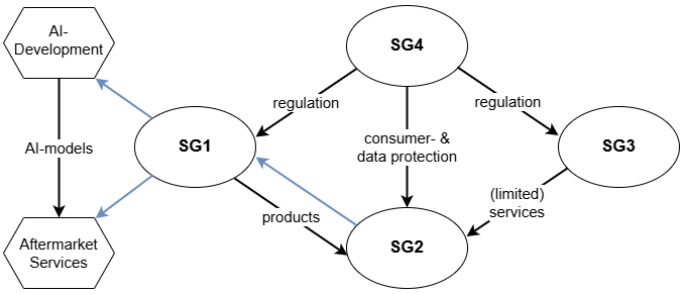


Figure 1. **Relationship diagram of the actors in the Smart Home and IoT** Legend: SG1 - Gatekeepers; SG2 - Users; SG3 - Aftermarket service providers; SG4 - Legislators and institutions; Data relations in blue.

SG1: Gatekeeper Technology companies such as Amazon and Google dominate the market by manufacturing devices and utilising data. Their focus is on increasing profits, often at the expense of data protection and despite strict regulations such as the GDPR [30][31]. As gatekeepers, they control the data that is created on devices produced by them. Therefore they have a significant influence on the market and competition [32]. Gatekeepers typically possess both privileged access to user data and the technical expertise to extract economic and strategic value from it.

SG2: Users Consumers benefit from automation and innovation, but are also the main source of data. Entertainment Systems, such as Smart TVs are the most popular among users, while building security and automation solutions are less common [33]. User priorities are security and convenience [34]. Customers and private users have limited technical capabilities and often only partial access to the data they generate.

SG3: Aftermarket service providers Companies offering repair services and cloud providers are heavily dependent on gatekeepers as they often lack access to critical data. They may have the technical know-how but face barriers in data access due to platform control and interoperability issues. This hampers innovation and fair competition [35].

SG4: Legislators and institutions Legislation, particularly at EU level, is aimed at data protection, consumer protection and a fair data market. Data from Smart Home systems could also be used in crises, such as pandemics or natural disasters [36].

III. Challenges of the Data Economy in the Smart Home

The data economy faces significant economic and legal hurdles, particularly in the Smart Home sector. A central problem are the gatekeepers - powerful technology companies that primarily operate outside the EU and exert considerable influence on the global flow of data. Their dominance makes it difficult to enforce European data protection regulations [37] and manifests itself in various power asymmetries vis-à-vis their customers. The trade in personal data and the data protection-compliant processing of this data pose further challenges, which can result in financial losses that threaten the existence of small and medium-sized companies in particular if the applicable regulations are not observed.

A. Data protection issues in the Smart Home

The data generated by Smart Home devices often contains sensitive and sensitive information about users and their everyday habits [6][38]. IoT technologies are expected to have a significant impact on the healthcare sector in particular, which further emphasises the sensitivity of this data [14]. Manufacturers of Smart Home devices and services have a significant influence on what data is collected and shared, while consumers are often insufficiently informed about the scope and storage of this data [39]. Without technical expertise, it is almost impossible for consumers to verify the accuracy of the specified data processing modalities and their control over the data they generate is severely limited [38]. In many cases, the only option available to users with data protection concerns is to opt out of the product or service.

Intelligent voice assistants are a particularly clear example of the lack of transparency in data processing. These systems require permanent monitoring of the acoustic environment in order to be able to react to trigger phrases [39]. Although data is only transmitted after a keyword has been recognised, voice assistants can be activated unintentionally, e.g. by similar-sounding phrases, whereby data is transmitted without the explicit request of the user [40]. Users are often left in the dark about the scope of the data collected, as detailed information can only be requested on their own initiative and in accordance with the applicable data protection laws, such as the GDPR. In addition, this data is usually not processed locally, but decentralised on the providers' cloud servers [41], which severely restricts the user's control over the transfer and processing of data.

To better understand the regulatory challenges in Smart Home environments, it is required to distinguish between different types of data and their sources. Table II classifies Smart Home data along two dimensions: the nature of the data (personal vs. non-personal) and its origin (individual users or shared use).

This classification in Table II highlights the complexity of data governance in multi-user settings, where personal and non-personal data often coexist and overlap, raising important questions about ownership and access rights.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE II
CLASSIFICATION AND EXAMPLES OF SMART HOME DATA BY USER AND TYPE

Data Type	User 1	User 2	Shared
Personal Data	voice assistant queries, health data	TV preferences, fitness data	shared calendar, living room camera footage
Non-Personal Data	generic device usage statistics (e.g., light switches)	app update logs, battery charging cycles	energy consumption, network diagnostics

The decision on how to handle the collected data often lies with the manufacturers, while the users, despite legal requirements such as [7, Art. 12-14], has no direct insight into or control over access to their data. In many cases, this data is sold to third parties or used by the provider to develop new services, often without the express consent of the user [39]. Even after the use of Smart Home devices, many providers retain the collected data, which increases the risk of future misuse or disclosure through security incidents [42].

Consumers are also dependent on manufacturers and service providers in their ability to protect their data, as they store the data in cloud systems [41][43]. Data protection-friendly functions, such as encryption or multi-factor authentication are often missing and can only be implemented by the manufacturer [39]. Particularly in connection with identifying data, such as payment or geolocation data, which by its very nature can be assigned to individuals, the security of this personal data depends largely on the security precautions taken by the cloud provider. In the event of a data leak, serious consequences, such as identity theft or financial damage can occur [44]. For companies that rely on networked devices, data protection incidents can also lead to a loss of sensitive business secrets, which harbours considerable economic risks [45].

There is a further risk in the second-hand trade for IoT and Smart Home devices. Due to the frequent lack of user interfaces on embedded devices, resetting used devices to factory settings is difficult and can result in the previous owner's personal data remaining on the device [46]. The new owner could unintentionally or maliciously gain access to this data or use functions that are linked to the previous owner's account.

B. Power asymmetries between consumers and service providers

In the course of the increasing networking of household appliances and the data-driven economy, power asymmetries between consumers and service providers (specifically SG1 'gatekeepers') are becoming ever more apparent. These result not only from the technical complexity, but also from the legal and economic conditions, which restrict consumers' scope for action and make access to the data they generate more difficult.

A central feature of these asymmetries are non-negotiable user agreements dictated by the provider. Particularly in the area of Smart Home technologies, providers impose opaque contractual terms and conditions [38], which can usually only be accepted by accepting or waiving the service. These often contain clauses that grant extensive rights of use to personal data or severely limit the provider's liability [47]. The exact scope of data use often remains opaque, which increases consumer mistrust [38]. The lack of transparency about data processing and the invisibility of processes that take place in remote data centres [41] further increase this scepticism. Access to and management of personal data often takes place via complex, less user-friendly platforms [48], which makes it difficult for many users to exercise their rights under the GDPR.

A particularly clear example of this power asymmetry is Amazon's Alexa voice assistant system. Here, the consumer has little control over the extent of data usage, as these processes are decentralised and hidden [49]. Users are dependent on a continuous connection to the cloud in order to use the service [50]. This increases dependency on the provider and makes it difficult to switch to alternative providers.

In addition, the lack of interoperability of cloud services makes it difficult to switch between different providers, as many systems are proprietary [51]. The repair options for Smart Home devices are also limited, as often only the manufacturer can carry out repairs [52].

These structural imbalances are leading to a loss of trust among consumers, who are increasingly trying to remove their data from the providers' streams. Data protection-oriented technologies such as Brave or the Tor browser, as well as specialised 'data removal services', are therefore gaining in importance [53]. This loss of trust, especially with regard to the technical security of networked devices, could damage the Smart Home sector in the long term as the value of the data collected decreases [33].

IV. EU DATA ACT AND GDPR IN THE SMART HOME

In this section, two central and widely applicable EU regulations with great relevance for the Smart Home and the data generated therein are analysed: the GDPR and the EU Data Act. These regulations aim to regulate the handling of personal and machine-generated data, strengthen consumer rights and address the issues analysed in Section II, such as data protection problems and power asymmetries between service providers and users.

Together with other legislative measures, for instance the EU Digital Markets Act, the EU Digital Services Act and the EU Data Governance Act, they form the basis of the European Commission's digital strategy.

A. Overview of the EU Data Act

The EU Data Act (EU Regulation 2023/2854) is a central element of the European data strategy and regulates access to and use of the generated data collected by digital technologies such as Smart Home devices and cloud computing. The regulation was adopted in November 2023, has been in force since the beginning of 2024 and will be applicable law from September 2025 [54]. The aim is to ensure fair access to

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

data and a fair distribution of data between different market players [5].

Manufacturers and service providers (data owners) are obliged to grant users prompt access to their generated data [6, Art. 3 para. 1]. This also includes the right to pass the data on to third-party providers who can use it to develop innovative products and services [6, Art. 5]. In the Smart Home sector in particular, this should help to ensure that not only large companies benefit from the data economy, but also smaller players [6, Recital 30 & 32].

The processing of personal data remains subject to the provisions of the GDPR, which takes precedence in the event of conflicts [6, Art. 1 para. 5]. The regulation is intended to give users more control over their data, as many consumers or companies do not have the resources to utilise the full economic value of their data themselves [6, Recital 3 & 19 & 40]. One example of implementation is the management of data by virtual assistants. These collect information about Smart Home users, for example to control heating or lighting. The EU Data Act enables users to manage this data and pass it on to third-party providers, which could give rise to new smart assistance systems [6, Recital 22].

In order to curb the market dominance of large platforms, gatekeeper companies defined under the EU Digital Markets Act may not use data that does not originate from their own devices [6, Art. 5 para. 3]. This is intended to prevent excessive concentration of market power [55][56]. In addition, the EU Data Act is intended to help better manage such public emergencies as pandemics or cyberattacks. In such cases, public authorities can request relevant data, whereby the modalities are clearly defined and companies may be compensated for providing it [6, Art. 14 & 17].

Figure 1 can thus be extended to Figure 2.

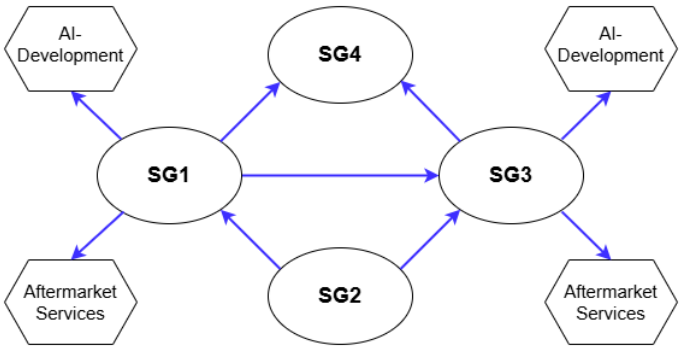


Figure 2. **In terms of the EU Data Act, modified relationship diagram of the actors in the Smart Home and IoT area from Figure 1**
Legend: SG1 - Gatekeepers; SG2 - Users; SG3 - Aftermarket service providers; SG4 - Legislators and institutions; Data relations in blue.

With the EU Data Act becoming applicable law, data that was previously exclusively accessible to gatekeepers (SG1) will be available to all authorised interested parties (see Figure 2). This enables European companies (SG3) to develop their own data-based services, while public authorities (SG4) can manage crises more efficiently.

However, the extended availability of data harbours security risks. Data owners are responsible for securing access and data, but can use modern security measures, for example encryption or smart contracts [6, Art. 3 para. 1, Art. 11 para. 1]. Users and recipients may only adapt these measures with the consent of the data controller without being restricted in their use of data.

Another key issue is cloud switching: users should be able to switch more easily between cloud providers, such as AWS or Microsoft Azure. Providers must support customers when switching [6, Art. 25 para. 2a], ensure business continuity and, from 2027, no longer charge fees for the move [6, Art. 29]. This incentivises structured, machine-readable data formats and technologies for smooth data migration [6, Recital 78 & 84].

Smart contracts are proposed as possible interfaces for automated data transfers to enable the secure and traceable execution of agreements [6, Art. 2 para. 39]. The EU Data Act requires high security standards, protection against manipulation and audit-proof archiving of the generated data [6, Art. 36]. Providers must regularly submit declarations of conformity to prove compliance with the regulations.

The regulation also emphasises security in critical infrastructures, for example energy and water supply [6, Recital 14]. In addition to the GDPR, security checks and measures such as pentests and encryption are also to be performed and implemented for non-personal data [6, Recital 102].

For implementation, national authorities are to be appointed or established to impose effective sanctions in the event of violations [6, Art. 37 & 40]. If several authorities are involved, a data coordinator will be appointed for coordination, while GDPR supervisory authorities will remain responsible for personal data [6, Art. 37 para. 3].

The EU Data Act addresses the asymmetric market power in the IoT sector in favour of users and strengthens competition [57, p. 26]. However, there are challenges: User consent remains a weak point, and interactions with the GDPR are still unclear [57, pp. 23]. Further legal and technical coordination is required in order to realise the full benefits of the EU Data Act.

B. Overview of the GDPR

The General Data Protection Regulation (GDPR) has been in force as binding law in the EU since 2018 and aims to ensure the protection of personal data [7]. It was introduced in response to increasing digitalisation and the growing amount of data collected [58]. Large tech companies in particular benefit from analysing and using such data [7, Recital 6].

A clear legal framework has been created to regulate data processing, which obliges companies to meet high standards and provides for sanctions in the event of violations [7, Recital 7]. In practice, however, there are difficulties with enforcement, particularly in cross-border cases [37].

The GDPR defines key terms such as *personal data*, *processing* and *profiling* [7, Art 4] and sets out binding principles for the handling of personal data. These include lawfulness,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

purpose limitation, data minimisation and accountability [7, Art. 5].

Data subjects have extensive rights, including rights to information, access and erasure [7, Art. 13-17]. Data controllers must ensure that these rights are respected and are obliged to appoint data protection officers and report data breaches within 72 hours [7, Art. 24 & 33].

An evaluation of the GDPR from 2024 emphasises the central role of data protection officers, especially for small and medium-sized enterprises [59, p. 3]. In addition, the regulation should be made less bureaucratic to enable more efficient and risk-based implementation [59, pp. 4]. Digital service providers should also be more closely integrated into the obligations of the GDPR in order to improve users' control over their data [59, pp. 5].

C. Comparison and conflicts between the two regulations

The GDPR and the EU Data Act have different priorities, but are not fundamentally contradictory. While the GDPR prioritises the protection of personal data and consumer protection, the EU Data Act focuses on promoting a data-driven economy and facilitating access to non-personal data [60]. Both sets of regulations share the goal of supporting the free movement of data within the EU by striving for a balance between data protection and the commercial use of data.

Both the GDPR and the EU Data Act contain provisions for crisis situations: While the EU Data Act permits the provision of relevant data, the GDPR authorises public bodies to process personal data in certain cases [7, Art. 2 para. 2d][6, Art. pp. 14]. These regulations can be helpful in the context of pandemics or disaster management, for example, but raise questions in the area of law enforcement and anti-terrorism measures. The centralised availability of data via data traders could also open up new opportunities for fighting crime, the effects of which need to be investigated further.

A central area of tension arises when networked devices generate personal data, as the EU Data Act makes it clear that it must not affect the GDPR and takes second place to it in the event of a conflict [6, Recital 34][6, Art. 1 para. 5]. This becomes particularly problematic in multi-user scenarios: A single user could share or sell data that also affects other people without their consent. This would be a violation of the GDPR [7, Art. 13]. At the same time, the EU Data Act obliges providers to provide non-personal data [6, Art. 3], which creates a legal dilemma: A refusal could violate the EU Data Act, a disclosure could violate the GDPR.

The distinction between personal and non-personal data poses a further challenge, especially in the case of mixed data sets. The EU Data Act requires a clear classification in order to provide commercially usable data, while the GDPR comprehensively protects personal data [6, Art. 3, 7, Art. 18]. It is particularly critical that originally non-personal data can become personal information through correlation or analytical procedures [61, p. 6 & 16][62]. This could lead to data protection provisions being circumvented through clever contractual constructions.

The practical implementation of both sets of regulations also poses challenges for companies. Small and medium-sized enterprises in particular are confronted with considerable bureaucratic effort due to the parallel requirements of the GDPR and the EU Data Act, which ties up resources and can inhibit innovation processes [59][60].

Approaches such as anonymisation or selective data sharing are being discussed to resolve these conflicts, although their technical feasibility and effectiveness remain questionable. A more precise regulation on the separation of personal and non-personal data as well as a clear legal handling of multi-user scenarios are necessary in order to make the coexistence of both sets of rules practicable [6, Recital 7].

D. Relevance for the Smart Home

Smart home devices, as part of the IoT, collect a lot of data in the home environment and are explicitly mentioned in the EU Data Act [6, Recital 23]. Currently, however, these systems are often limited by incompatibilities, while users find it difficult to access their own data [35][63]. The EU Data Act is intended to counteract this by promoting better data accessibility and interoperability between manufacturers [6, Recital 32]. This not only enables personalised services, but also facilitates the repair of defective devices through improved data access [6, Recital 32].

As considerable amounts of data are generated in the Smart Home, these are not only of economic interest to users, but also to companies. The EU Data Act obliges providers to make this data available - both to consumers and to third parties, including competitors [6, Recital 39].

The data collected affects many areas of life, from health to entertainment [64], and offers both individual and economic benefits [6, Recital 64]. At the same time, they are often personal [38] and are therefore subject to the GDPR, which creates data protection risks, especially through the detailed recording of user behaviour [38].

Heino et al. [38] describes four central data protection problems in the Smart Home:

(1) Unclear legal scope of application,
(2) Lack of transparency in data processing,
(3) preset data collection with opt-out instead of opt-in and
(4) uncertainties regarding retention periods. Added to this is the principle of data minimisation, which conflicts with the usefulness of many devices, as more data often means better functionality [61, p. 15][65, p. 3].

Following our analysis, we would expand this list based on the differences between the kinds of data generated in the smart home as described in Table II:

(5) Blurred boundaries between personal and non-personal data, especially in shared or mixed-use contexts,
(6) Ambiguity in attributing data to specific individuals in multi-user environments,
(7) Unclear responsibilities for data governance when data is co-generated or shared across devices and users,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

(8) Conflicts between user rights under the Data Act (e.g., data portability) and the privacy rights of other users under the GDPR.

Since the GDPR came into force, users in countries with GDPR-compliant legislature are increasingly aware of data security risks and have a higher perceived level of control [66]. The multi-user operation of Smart Home devices makes GDPR compliance more difficult. In households, several people share devices, which makes it challenging to clearly assign and control personal data. A flexible solution is needed to combine data protection and usability.

V. Enforcement of the GDPR: Trends and Implications

Since the GDPR entered into force in 2018 [7, Art. 99], supervisory authorities across the EU have regularly imposed fines for infringements.

By September 2024, total fines amounted to roughly €5 billion across more than 2,000 cases [67], demonstrating active enforcement and significant financial consequences. The sharp increases in July 2021 and May 2023 are linked to rulings against *Amazon* (€750 million) [68] and *Meta Platforms* (€1.2 billion) [69].

The most striking case remains *Meta Platforms*, fined more than €2 billion since 2018 [67]. These sanctions primarily concern the lawfulness of processing [7, Art. 6], data processing principles [7, Art. 5], and insufficient technical and organizational safeguards [7, Art. 32]. Yet, with 2023 revenues of €134.9 billion [70], these fines account for less than 1.5 % of Meta's annual turnover, raising doubts about their deterrent effect.

Beyond Meta, fines often stem from violations of lawful processing [7, Art. 6], fundamental processing principles such as integrity and confidentiality [7, Art. 5f & 32], and failure to respect data subject rights [7, Art. 15ff.]. Smaller cases frequently involve non-compliance with access or erasure obligations, with fines ranging from €100 to €10,000 [67]. For SMEs, the administrative burden of meeting these requirements without dedicated compliance units poses a significant challenge [71].

The effectiveness of the GDPR as a deterrent remains contested [31][59][60][68]. Large repeat offenders like Meta call into question the impact of fines, while SMEs often comply more carefully, as penalties can threaten their existence. Thus, the deterrent effect of sanctions appears uneven across company sizes and sectors.

Regardless of its ultimate efficacy, the history of imposed fines underlines the prominent role of sanctions in enforcing regulatory standards, especially in the domain of data protection. Given the parallels in scope and ambition, similar enforcement intensity can be expected under the Data Act. As discussed in Section IV, the combination of GDPR and Data Act obligations creates particularly complex challenges in multi-user environments such as Smart Homes. Proactive compliance strategies are therefore essential to safeguard both user rights and companies from future sanctions.

VI. Case Study: Bosch Smart Home Controller

An illustrative case study is provided by Bosch's Smart Home Controller, which is documented in an openly available GitHub repository under https://github.com/BoschSmartHome/bosch-shc-api-docs. The system enables the centralized management of supported Smart Home devices and the data they generate. It serves as a useful example of how Smart Home ecosystems can be structured in a way that aligns with the requirements of the EU Data Act. In particular, Bosch provides standardized interfaces for accessing device data and facilitates interoperability with third-party services, which reflects the Act's emphasis on data availability and portability.

The Bosch Smart Home Controller exposes a local REST API, allowing developers and integrators access to a comprehensive set of functionalities: device status queries, event subscription, automation rules management, and retrieval of historical logs. API communication is encrypted with TLS 1.2, and authorized clients must register via secure key exchange, setting a technical baseline for secure integration. The system also supports modern connectivity protocols such as Zigbee and Matter, enabling compatibility with a wide range of third-party devices and aligning with broader interoperability goals in the smart home market.

However, a closer analysis reveals some shortcomings with regard to user-level data governance. The controller does not implement device-level access control for different household members. Instead, the designated administrator account has unrestricted access to all data generated within the Smart Home environment, including personal data of other users. While this design significantly increases usability and system maintenance, it stands in direct tension with the principle of user-specific rights over personal data under the GDPR[7, Art. 12-23]. The absence of mechanisms to separate or shield multi-user data illustrates a critical challenge for ensuring compliance in real-world Smart Home infrastructures.

Additional concerns relate to the granularity and transparency of data handling. Although Bosch provides well-documented APIs for data access and integration, the system lacks fine-grained role management and user-specific permission settings. Logs of who accessed or exported which data are not visible to individual users, thereby limiting transparency and accountability. Furthermore, while interoperability with external services is supported, safeguards for preventing the uncontrolled sharing of mixed datasets remain underdeveloped. The availability of technical integration points is an important step toward data portability, but the absence of robust audit mechanisms and user-centered access controls highlights a gap between the formal requirements of EU data governance frameworks and practical implementation in commercial smart home solutions.

These observations underline the difficulty of bridging the gap between competing regulatory aspirations and the current state of technical implementation in Smart Home platforms, highlighting the need for privacy-by-design features that operationalize legal requirements.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

VII. Conclusion and Future work

The Smart Home has been identified as an increasingly relevant component of the data economy, in which the residents of a networked living space generate a large amount of data through their use of intelligent devices. This data offers significant potential for new services, such as personalised energy management systems or health-applications, but also creates risks. These include data protection problems, security gaps and the possibility of commercial exploitation of user data by third parties.

The Smart Home sector is characterised by a wide range of actors, including manufacturers of IoT devices, service providers, public authorities and consumers themselves. These players often pursue divergent interests: While providers primarily aim to monetise user data, consumers demand stronger data protection measures and easy ways to control their data. The power asymmetries between large technology providers and their customers or small to medium-sized enterprises were identified as particularly problematic, as they hinder innovation and competition in the sector.

The forthcoming applicability of the EU Data Act and the existing requirements of the GDPR are having a significant impact on the Smart Home sector. The EU Data Act addresses the current power asymmetry by facilitating data access for data-generating customers and third-party providers, thereby promoting competition and innovation.

The GDPR, on the other hand, emphasises the protection of personal data and defines strict requirements for its processing. In combination, these two regulations create a complex legal framework and therefore pose considerable challenges. Conflicting objectives arise, particularly in the case of mixed data sets that contain both personal and non-personal information of several users: while the EU Data Act requires the release of non-personal data, the GDPR demands strict protective measures for personal data. This leads to uncertainties as to how the two regulations can be harmonised without violating data protection regulations and fearing sanctions. If such uncertainties are not adequately addressed, the Data Act may have the opposite of its intended effect and hinder innovation through diminished customer acceptance and reduced trust in data-driven services.

Our analysis of GDPR enforcement demonstrates that sanctions play a crucial role in ensuring compliance, though their effectiveness varies significantly between large technology firms and smaller companies. At the same time, the Bosch Smart Home Controller case study illustrates the practical challenges of implementing regulatory requirements in real-world systems, where multi-user environments and insufficient access controls create gaps between legal obligations and technical practice. Taken together, these findings underline both the necessity of effective enforcement mechanisms and the importance of privacy-by-design solutions if the goals of the GDPR and the EU Data Act are to be realized in the Smart Home sector.

The Bosch Smart Home Controller case study illustrates both achievements and ongoing challenges in aligning smart home systems with EU Data Act and GDPR requirements. While Bosch demonstrates progress in data accessibility and interoperability through standardized APIs, the system lacks the fine-grained user controls and transparency mechanisms necessary for GDPR compliance, particularly in multi-user households. Specifically, the centralized administrator model grants unrestricted access to all household data without role-based separation, creating tension with GDPR principles of individual user rights. These findings underscore that effective compliance requires privacy-by-design features embedding user-specific data governance at the infrastructure level, not merely API documentation and interoperability support.

In view of regulatory developments and the increasing spread of Smart Home technologies, several relevant research questions arise. One key issue is the practical implementation of the EU Data Act in the area of conflict with the GDPR, particularly with regard to the separation of personal and non-personal data within mixed data sets. There is a need for further clarification here to ensure that both data protection requirements and economic interests are adequately taken into account.

Another research approach is the development of technical solutions for more data protection and data sovereignty in the Smart Home. Approaches such as federated learning or edge computing could help to make data processing more decentralised in order to reduce security risks and power asymmetries. There is also a need for further research into how users can obtain intuitive and effective control mechanisms over their data without compromising the user-friendliness of Smart Home systems.

Finally, the economic and social impact of the new regulations must also be analysed. It remains to be seen to what extent the EU Data Act will actually promote competition in the Smart Home market or whether new challenges will arise due to regulatory uncertainties. Further research could focus on what adjustments the industry needs to make in order to fulfil the legal requirements and develop innovative, data protection-compliant business models.

VIII. Acknowledgment

This paper was funded by the European Union and the Free State of Saxony (Germany).

Kofinanziert von der Europäischen Union

Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des vom Sächsischen Landtag beschlossenen Haushaltes.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

References

[1] P. Seidel, F. Fischer, and D. Labudde. "Consequences of the EU Data Act and the EU General Data Protection Regulation for the Modern Smart Home Data Economy". In: *Proceedings of the 19th International Conference on Digital Society (ICDS 2025)*. IARIA. 2025. ISBN: 978-1-68558-267-8. URL: https://www.thinkmind.org/articles/icds_2025_1_40_10025.pdf.

[2] M. Szczepanski. *Is data the new oil?* accessed: 09.04.2025. Jan. 2020. URL: https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/646117/EPRS_BRI(2020)646117_EN.pdf.

[3] M. Chui et al. *The economic potential of generative AI - the next productivity frontier*. accessed: 09.04.2025. 2023. URL: http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/14313/1/The-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf.

[4] Publications Office of the European Union. *Data Act — Factsheet*. accessed: 09.04.2025. 2022. DOI: 10.2775/636433. URL: https://digital-strategy.ec.europa.eu/en/library/data-act-factsheet.

[5] Council of the EU. *Data Act: Council adopts new law on fair access to and use of data*. accessed: 09.04.2025. Nov. 2024. URL: https://www.consilium.europa.eu/en/press/press-releases/2023/11/27/data-act-council-adopts-new-law-on-fair-access-to-and-use-of-data/.

[6] European Parliament and Council of the EU. "Regulation (EU) 2023/2854 of the European Parliament and of the Council". of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). In: *Official Journal of the European Union* (Dec. 2023). accessed: 09.04.2025. URL: https://eur-lex.europa.eu/eli/reg/2023/2854/oj.

[7] European Parliament and Council of the EU. "Regulation (EU) 2016/679 of the European Parliament and of the Council". of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). In: *Official Journal of the European Union* (May 2016). accessed: 09.04.2025. URL: https://data.europa.eu/eli/reg/2016/679/oj.

[8] D. Kerr, N. Robins-Early, and J. Bhuiyan. "'Slap on the wrist': critics decry weak penalties on Google after landmark monopoly trial". In: *The Guardian* (Sept. 2025). URL: https://www.theguardian.com/technology/2025/sep/03/google-monopoly-case-ruling.

[9] S. Heath. *Embedded Systems Design*. accessed: 09.04.2025. 2003. ISBN: 0-7506-5546-1. URL: https://books.google.de/books?hl=en&lr=&id=BjNZXwH7HlkC&oi=fnd&pg=PP1&dq=embedded+systems&ots=xk1jG-BMkP&sig=_hYgwJ2cR_bfy4DNTNK3M3PCsN4&redir_esc=y#v=onepage&q=embedded%20systems&f=false.

[10] Federal Network Agency. *Federal Network Agency - Internet of Things*. accessed: 10.04.2025. Oct. 2024. URL: https://web.archive.org/web/20250328092032/https://www.bundesnetzagentur.de/DE/Fachthemen/Digitalisierung/Internet/IoT/start.html.

[11] Ibm. *Internet of Things*. May 2024. URL: https://www.ibm.com/de-de/topics/internet-of-things.

[12] B. L. Risteska Stojkoska and K. V. Trivodaliev. "A review of Internet of Things for smart home: Challenges and solutions". In: *Journal of Cleaner Production* 140 (2017). ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2016.10.006. URL: https://www.sciencedirect.com/science/article/pii/S095965261631589X.

[13] L. Ferreira, T. Oliveira, and C. Neves. "Consumer's intention to use and recommend smart home technologies: The role of environmental awareness". In: *Energy* 263 (2023), p. 1. ISSN: 0360-5442. DOI: 10.1016/j.energy.2022.125814. URL: https://www.sciencedirect.com/science/article/pii/S0360544222027001.

[14] A. Al-Fuqaha et al. "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications". In: *IEEE Communications Surveys & Tutorials* 17.4 (2015), pp. 2347–2376. DOI: 10.1109/COMST.2015.2444095.

[15] B.-O. Bat-Erdene and J. L. Saver. "Automatic Acute Stroke Symptom Detection and Emergency Medical Systems Alerting by Mobile Health Technologies: A Review". In: *Journal of Stroke and Cerebrovascular Diseases* 30.7 (2021), p. 105826. ISSN: 1532-8511. DOI: 10.1016/j.jstrokecerebrovasdis.2021.105826. URL: https://www.strokejournal.org/article/S1052-3057(21)00229-9/fulltext.

[16] NHLBI. *Novel sensor can detect a heart attack in just minutes*. accessed: 09.04.2025. Oct. 2021. URL: https://www.nhlbi.nih.gov/news/2021/novel-sensor-can-detect-heart-attack-just-minutes.

[17] M. Zehnder et al. "Energy saving in smart homes based on consumer behavior: A case study". In: *2015 IEEE First International Smart Cities Conference (ISC2)*. 2015. DOI: 10.1109/ISC2.2015.7366231.

[18] A. Anvari-Moghaddam, H. Monsef, and A. Rahimi-Kian. "Optimal Smart Home Energy Management Considering Energy Saving and a Comfortable Lifestyle". In: *IEEE Transactions on Smart Grid* 6.1 (2015), pp. 324–332. DOI: 10.1109/TSG.2014.2349352. URL: https://ieeexplore.ieee.org/abstract/document/6895131.

[19] Telecom. *SmartHome Ideas: Heating Control*. accessed: 09.04.2025. Aug. 2024. URL: https://web.archive.org/web/20240915162832/https://www.smarthome.de/ideen/smarte-heizungssteuerung.

[20] M. Bräuer. *ABUS Safety Study 2023*. 2023.

[21] J. Dahmen et al. "Smart Secure Homes: A Survey of Smart Home Technologies that Sense, Assess, and Respond to Security Threats". In: *Journal of reliable intelligent environments* 3.2 (2017), p. 1. DOI: 10.1007/

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

s40860-017-0035-0. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5616189/.

[22] L. Miller. *Integrating Smart Building Predictive Maintenance to Your System*. Ed. by L. Miller. accessed: 09.04.2025. Oct. 2024. URL: https://web.archive.org/web/20240814031401/https://www.buildingsiot.com/blog/integrating-smart-building-predictive-maintenance-to-your-system-bd.

[23] P. L. Austin. *What Will Smart Homes Look Like 10 Years From Now?* accessed: 10.04.2025. July 2019. URL: https://time.com/5634791/smart-homes-future/.

[24] X. Guo et al. "Review on the Application of Artificial Intelligence in Smart Homes". In: *Smart Cities* 2.3 (2019), p. 1. ISSN: 2624-6511. DOI: 10.3390/smartcities2030025. URL: https://www.mdpi.com/2624-6511/2/3/25.

[25] L. Budach et al. *The Effects of Data Quality on Machine Learning Performance*. 2022. URL: https://api.semanticscholar.org/CorpusID:251223513.

[26] Y. Chen. "IoT, cloud, big data and AI in interdisciplinary domains". In: *Simulation Modelling Practice and Theory* 102 (2020), p. 1. ISSN: 1569-190X. DOI: 10.1016/j.simpat.2020.102070. URL: https://www.sciencedirect.com/science/article/pii/S1569190X20300083.

[27] Council of the EU. *Data Act: Council adopts new law on fair access to and use of data*. accessed: 10.04.2025. Oct. 2024. URL: https://commission.europa.eu/strategy-and-policy/coronavirus-response/travel-during-coronavirus-pandemic/contact-tracing-and-warning-apps-during-covid-19_de.

[28] R. K. Institute. *Digitally interrupt chains of infection with the Corona-Warn-App*. accessed: 11.04.2025. May 2024. URL: https://www.rki.de/DE/Themen/Infektionskrankheiten/Infektionskrankheiten-A-Z/C/COVID-19-Pandemie/CoronaWarnApp/Warn_App.html.

[29] D. P. Conference. *Use of digital contact tracing services for event, facility, restaurant and business visits to prevent the spread of Covid-19: Guidance from the Conference of Independent Federal and State Data Protection Supervisory Authorities*. Apr. 2021.

[30] EU Commission - Representation in Germany. *Fair digital markets: gatekeepers must comply with all DMA rules starting today*. accessed: 10.04.2025. Mar. 2024. URL: https://germany.representation.ec.europa.eu/news/faire-digitale-markte-torwachter-mussen-ab-heute-alle-dma-regeln-einhalten-2024-03-07_de.

[31] European Commission. *COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL: Second report on the application of the General Data Protection Regulation (GDPR)*. accessed: 10.04.2025. July 2024. URL: https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52024DC0357.

[32] European Parliament and Council of the EU. "REGULATION (EU) 2022/1925 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL". of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). In: *Official Journal of the European Union* (Dec. 2023). accessed: 09.04.2025. URL: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32022R1925.

[33] Deloitte. *Consumer IoT fact check: The Internet of Things in the everyday lives of German consumers*. accessed: 09.04.2025. 2021. URL: https://web.archive.org/web/20231214185700/https://www.deloitte.com/content/dam/Deloitte/de/Documents/technology-media-telecommunications/Consumer_IoT_2021_Deloitte.pdf.

[34] B. K. Sovacool, D. D. Furszyfer Del Rio. *Smart home technologies in Europe: A critical review of concepts, benefits, risks and policies*. accessed: 10.04.2025. 2022. URL: https://www.sciencedirect.com/science/article/pii/S1364032119308688.

[35] European Commission. *Data Act – Questions and Answers*. June 2023. URL: https://ec.europa.eu/commission/presscorner/detail/en/qanda_22_1114.

[36] European Commission. *European data strategy*. accessed: 10.04.2025. Nov. 2024. URL: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en.

[37] EU Commission - Representation in Germany. *General Data Protection Regulation: EU Commission wants to improve enforcement*. accessed: 10.04.2025. July 2023. URL: https://germany.representation.ec.europa.eu/news/datenschutzgrundverordnung-eu-kommission-will-durchsetzung-verbessern-2023-07-04_de.

[38] T. Heino, S. Rauti, and R. Carlsson. "An assessment of privacy policies for smart home devices". In: *Proceedings of the 24th International Conference on Computer Systems and Technologies (CompSysTech '23)*. Ed. by T. Vassilev and R. Trifonov. ACM Other conferences. New York, NY, United States: Association for Computing Machinery, 2023, p. 1. ISBN: 979-8-4007-0047-7. DOI: 10.1145/3606305.3606332. URL: https://dl.acm.org/doi/fullHtml/10.1145/3606305.3606332.

[39] R. Herold. *Five Common Privacy Problems in an Era of Smart Devices*. accessed: 10.04.2025. Jan. 2020. URL: https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2020/five-common-privacy-problems-in-an-era-of-smart-devices.

[40] L. Schoenherr, M. Golla, T. Eisenhofer, J. Wiele, D. Kolossa, and T. Holz. *Unacceptable, where is my privacy?* accessed: 09.04.2025. Nov. 2021. URL: https://unacceptable-privacy.github.io/index.html.

[41] A. R. Biswas and R. Giaffreda. "IoT and cloud convergence: Opportunities and challenges". In: *Internet of Things WF-IoT*. 2014. DOI: 10.1109/WF-IoT.2014.6803194.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[42] N. Olsen. *Voice Assistants and Privacy Issues - Privacy Policies*. Ed. by N. Olsen. accessed: 10.04.2025. June 2022. URL: https://www.privacypolicies.com/blog/voice-assistants-privacy-issues/.

[43] Consumer advice center. *Smart Home - The intelligent Home*. accessed: 10.04.2025. Oct. 2024. URL: https://www.verbraucherzentrale.de/wissen/umwelt-haushalt/wohnen/smart-home-das-intelligente-zuhause-6882.

[44] Federal Office for Information Security. *Identity theft via data leaks and doxing*. URL: https://www.bsi.bund.de/dok/6692610.

[45] J. Thorpe-Smith. *Data Leaks: The Biggest Risks, Consequences, Causes & How to Prevent Them — Metomic*. Nov. 2024. URL: https://www.metomic.io/resource-centre/what-are-the-biggest-risks-of-data-leaks.

[46] dpa Service. *Securely networked?: how data protection works in the smart home*. accessed: 10.04.2025. May 2021. URL: https://www.zeit.de/news/2021-05/20/so-geht-datenschutz-im-smarthome.

[47] Dr. T. J. Heydn. *Advice on drafting contracts for the Internet of Things (IoT)*. accessed: 10.04.2025. 2021. URL: https://www.tcilaw.de/hinweise-zur-vertragsgestaltung-beim-internet-of-things-iot/.

[48] Consumer advice center NRW. *Data disclosure: How to find out what companies know about you*. accessed: 10.04.2025. 2023. URL: https://www.verbraucherzentrale.nrw/wissen/digitale-welt/datenschutz/datenauskunft-so-erfahren-sie-was-unternehmen-ueber-sie-wissen-44238.

[49] Amazon.de. *Amazon.de Privacy Notice*. accessed: 10.04.2025. 2024. URL: https://www.amazon.de/gp/help/customer/display.html/ref=footer_privacy?ie=UTF8&nodeId=GX7NJQ4ZB8MHFRNJ&language=en_GB.

[50] Amazon.de. *Alexa Terms of Use*. accessed: 10.04.2025. 2023. URL: https://www.amazon.de/gp/help/customer/display.html?nodeId=201809740.

[51] *What is AWS Database Migration Service? - AWS Database Migration Service*. accessed: 09.04.2025. 2024. URL: https://docs.aws.amazon.com/dms/latest/userguide/Welcome.html.

[52] European Commission. *Right to repair: Questions & Answers*. accessed: 10.04.2025. 2023. URL: https://ec.europa.eu/commission/presscorner/api/files/document/print/en/qanda_23_1795/QANDA_23_1795_EN.pdf.

[53] Q. Baterna. *6 Ways to Make It Harder for Data Brokers to Collect Your Data*. accessed: 10.04.2025. Nov. 2021. URL: https://www.makeuseof.com/ways-to-make-it-harder-for-data-brokers-collect-your-data/.

[54] Federal Ministry for Digital and Transport. *EU verabschiedet Data Act*. accessed: 10.04.2025. 2023. URL: https://web.archive.org/web/20250323125138/https://bmdv.bund.de/DE/Themen/Digitales/Digitale-Gesellschaft/EU-Data-Act/eu-data-act.html.

[55] European Commission. *Frequently Asked Questions - Data Act*. accessed: 10.04.2025. 2024. URL: https://digital-strategy.ec.europa.eu/de/library/commission-publishes-frequently-asked-questions-about-data-act.

[56] Dr. D. Pauly and Dr. A. Lohbeck. *EU: Data Act – How fair is the draft regulation for more fairness in the data economy?* accessed: 10.04.2025. 2022. URL: https://linklaters.de/insights/publikationen/tmt/datenschutz/2022/februar/eu-data-act.

[57] Prof. Dr. W. Kerber. *The EU "Data Act": A Critical Analysis*. accessed: 10.04.2025. 2022. URL: https://www.uni-marburg.de/de/fb02/professuren/vwl/wipol/prof-wolfgang-kerber/presentation/2022_03_21-kerber-pres-data-act.pdf.

[58] European Data Protection Supervisor. *History of the development of the GDPR*. accessed: 10.04.2025. Dec. 2024. URL: https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_de.

[59] Stiftung Datenschutz (Foundation for Data Protection), BDV e.V., and DIHK. *Data Act and GDPR: For more legal clarity on data access and use*. accessed: 10.04.2025. 2022. URL: https://dsgvo-2024.org/wp-content/uploads/2022/08/BvD_Stiftung-Datenschutz_DIHK_Positionspapier_Data-Act-und-DSGVO_v1.pdf.

[60] M. Goetz and Prof. Dr. B. P. Paal. *Between data use and data protection: The relationship between the GDPR, Data Act and Data Governance Act*. Dec. 2024. URL: https://stiftungdatenschutz.org/veranstaltungen/unsere-veranstaltungen-detailansicht/zwischen-datennutzung-und-datenschutz-438.

[61] S. Piasecki. "Expert perspectives on GDPR compliance in the context of smart homes and vulnerable persons". In: *Information & Communications Technology Law* 32.3 (2023), pp. 385–417. ISSN: 1360-0834. DOI: 10.1080/13600834.2023.2231326. URL: https://www.tandfonline.com/doi/full/10.1080/13600834.2023.2231326#d1e109.

[62] Harvard Business School Online. *Data Analytics Privacy Issues & How to Avoid Them*. accessed: 10.04.2025. 2015. URL: https://online.hbs.edu/blog/post/data-privacy-issues.

[63] K. Ahuja and M. Patel. *There's No Place Like A Connected Home: Perspectives on the connected consumer in a world of smart devices*. accessed: 10.04.2025. Nov. 2020. URL: https://www.mckinsey.com/spContent/connected_homes/index.html.

[64] J. Bugeja, A. Jacobsson, and P. Davidsson. "An Empirical Analysis of Smart Connected Home Data". In: *Internet of Things - ICIOT 2018*. Ed. by D. Georgakopoulos and L.-J. Zhang. Information Systems and Applications, incl. Internet/Web, and HCI. Cham: Springer International Publishing and Imprint: Springer, 2018, pp. 134–149. ISBN: 978-3-319-94370-1. DOI: 10.1007/978-3-319-94370-1_10. URL: https://link.springer.com/chapter/10.1007/978-3-319-94370-1_10.

[65] D. Bastos et al. *GDPR Privacy Implications for the Internet of Things*. 2018. URL: https://www.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

researchgate . net / publication / 331991225_GDPR_ Privacy_Implications_for_the_Internet_of_Things.

[66] V. Dahl and M. Österlin. *Impact of GDPR on Data Sharing Behavior of Smart Home Users*. Oct. 2020.

[67] enforcementtracker.com. *GDPR Enforcement Tracker - list of GDPR fines*. accessed: 02.12.2024. 2/12/2024. URL: https://www.enforcementtracker.com/?insights.

[68] M. Holzhofer. "746 Mio. Euro Bußgeld gegen Amazon Europe Core S.à r.l". In: *Compliance Essentials GmbH* (30/07/2021). accessed: 02.12.2024. URL: https://www. dsgvo - portal . de / bussgelder / dsgvo - bussgeld - gegen - amazon-europe-2021-07-30-LU-1448.php.

[69] European Data Protection Board. "Binding Decision 1/2023 on the dispute submitted by the Irish SA on data transfers by Meta Platforms Ireland Limited for its Facebook service (Art. 65 GDPR". In: (2023). accessed: 12.02.2024. URL: https://www.edpb.europa.eu/system/ files / 2023 - 05 / edpb_bindingdecision_202301_ie_sa_ facebooktransfers_en.pdf.

[70] META Platforms Inc. *Meta Reports Fourth Quarter and Full Year 2023 Results; Initiates Quarterly Dividend*. accessed: 09.04.2025. 1/02/2024. URL: https://investor. fb.com/investor-news/press-release-details/2024/Meta-Reports-Fourth-Quarter-and-Full-Year-2023-Results-Initiates-Quarterly-Dividend/default.aspx.

[71] C. Runte et al., eds. *Enforcement Tracker Report 2024*. accessed: 09.04.2025. 2024. URL: https://cms.law/en/ media/international/files/publications/publications/gdpr-enforcement-tracker-report-may-2024?v=5.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Defect Prevention Matrix Analysis by Generative AI

Shuichiro Yamamoto
Department of Information Technology
IPUT in Nagoya
4-27-1 Meieki Nakamura-ku, Nagoya, Japan
yamamoto.shuichiro@gmail.com

***Abstract*—To clarify the integrity of business processes, we propose a business process diagram that describes six aspects: input, output, acceptance conditions, resource conditions, exception conditions, and decision conditions. By separating exception conditions from output, defect prevention diagrams have the advantage of being able to detect defects and extract corresponding exception handling knowledge. However, methods for reviewing defect prevention diagrams have not been clear. In this paper, we define a process relationship matrix and present a review procedure to prevent business process defects. We also demonstrate through application experiments that the review procedure can be automated using generative AI and that six aspects of issues can be detected for each process. The main results of this paper are a business process review method using a process relationship matrix and a prompt template that automates this process using generative AI.**

Keywords- Business Process Completeness; Exception; Knowledge Management; Defect Prevention; Process Relationship Matrix; Generative AI.

I. INTRODUCTION

First, we note that this paper is an extension of the work described in [1]. In this paper, we demonstrate that the procedure proposed in the international conference paper can be automated by defining it as a prompt for generative AI, and we also add application examples to demonstrate its effectiveness. Furthermore, we clarify that the purpose of the proposed method is to detect defects in production and operational processes, not to analyze defects in product shape.

Business process modeling is necessary for corporate digital transformation. While notational methods for defining business models have been proposed to date, they pose the following problems.

If the purpose and conditions for achieving a business process cannot be clearly described, the goals that stakeholders must agree on are unclear, making it impossible to confirm the completeness of the business process.

The execution order of business processes alone not only makes the resources required for process execution and completion conditions unclear. Also, it is impossible to control priority when inputs from multiple preceding processes conflict. Furthermore, because exceptions that occur in the business process are not clearly defined, it is impossible to design appropriate responses.

For this reason, defect prevention diagrams have been proposed. Defect prevention diagrams address the above issues by describing six aspects of each business process: input, output, acceptance conditions, resource conditions, decision conditions, and exception conditions. Furthermore, the validity of a defect prevention diagram can be confirmed using a business process relationship matrix, which clearly defines the relationships between business processes.

This paper presents an analytical method using a defect prevention matrix to review business processes and a prompt template that automates the analysis using generative AI.

The rest of this paper is organized as follows. Section II describes related work. Section III describes the defect prevention matrix analysis. Section IV explains experimental case study. Section V presents discussions. The conclusion closes the article.

II. RELATED WORK

Related studies on Ji-Koutei-Kanketsu (JKK), Knowledge transfer, Business Process Modeling (BPM), Defect Prevention Analysis (DPA) and Functional Resonance Analysis Method (FRAM) are explained below.

A. Ji Koutei Kanketsu

In the production process, there is a misconception that local optimization is necessary, as long as one's own process is fine, and that unnecessary problems shall not be introduced to one's department. If a problem is discovered at the final stage of development, the design cannot be modified, or the basic structure of the product cannot be changed. Therefore, comprehensive product design and manufacturing are required throughout the entire production process. Ji-Koutei-Kanketsu (JKK) is a method that optimizes the entire production process, not just a specific process. The Japanese words Ji, Koutei, and Kanketsu [2] are self, process, and completion, respectively.

To introduce JKK, it is necessary to define not only business procedures that define the flow of work, but also requirements organization sheets that define business requirements. The requirements organization sheet consists of fields of the necessary items/information, business inputs, and business outputs for each business process. The necessary item and information field clarifies the input, tools,

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

methods, capabilities/authority, and reasons as conditions for the quality of the product. The input field describes the receiving criteria, such as when, where, and what. The output field describes where to sink, by when, and what to produce. The criteria field describes criteria for determining that "the output of the process is good."

JKK's production processes can also be seen as business processes. JKK clarifies the completeness conditions for each business process element. The requirement organization sheet is an essential feature of JKK.

B. Knowledge transfer

In order to transfer a company's experiential knowledge, it is necessary to clarify business processes. For this reason, methods for clarifying business processes have been proposed for knowledge transfer.

From a knowledge perspective, processes need to be defined to provide appropriate knowledge for tasks in an organization's operational business processes. In addition, knowledge must be extracted for the long-term growth, development, and competitiveness of companies. However, unless valuable knowledge within an organization is externalized or formalized, it cannot be used by other employees and disappears from the company. Therefore, Knowledge management shall be established using Business Process Modeling (BPM). Salvadorinhoa and Teixeira [3] pointed that BPM can not only help organizations improve their Industry 4.0 environment but also facilitate knowledge acquisition and distribution.

C. Business Process Modeling

Ore et al. [4] proposed a Self-managed organization based on Business Process Management. They showed a need for the business process management approach, which would manage the need for keeping critical business processes continuity and self-managed way of working of autonomous teams.

As long as the digitalization of business is promoted, business process documentation becomes vital for business process continuity. The digitalization re-constructs the traditional business processes into new digitalized business processes [5]. For example, the Digital Balanced Scorecard (DBSC) [6] consists of digital business processes.

There are many Business Model notations, including Business Process Models. Yamamoto [7] compared the representation capability of Business Model notations by defining fifteen key features of these notations with five interrogatives.

Leonard and Swap [8] defined deep smart as the expertise that allows experts to instantly grasp complex situations and make quick and wise decisions to deal with real problems. That is, deep smart is strong expertise formed by beliefs and social influences that can generate insights based on tacit knowledge grounded in direct experience. For example, in production process design, the problem is how to transfer defect investigation knowledge from experienced workers to beginners. An example of deep smart is the failure investigation knowledge that experienced engineers have. Leonard and Swap pointed out the importance of acquiring empirical knowledge through experimental learning. However, no concrete experimental learning method has been clarified. In addition, they have not clarified the knowledge representation of deep smarts. If deep smart cannot be expressed, it remains tacit knowledge, and deep smart knowledge transfer from experts to beginners is individual and difficult to spread horizontally.

As a technique for improving production processes in the manufacturing industry, Mono-Koto-Bun-seki (MKB) (in Japanese) has been proposed [9]. Mono, Koto, and Bun-seki mean Entity, Process, and Analysis, respectively. By treating objects such as materials and products as “entities” and the series of activities that make products from materials as "process," MKB can analyze the production process, discover waste, and optimize it.

Yamamoto and Fujimoto [10] proposed the Production Knowledge Chart (PKC) that expresses the production process to acquire the empirical knowledge necessary for investigating defects in manufacturing processes.

Object Process Methodology (OPM) proposed by Dori includes Object and Process [11][12]. For example, the aircraft design OPM has a Stakeholder Needs Set, Assumptions and Constraints Sets, and Requirements as Objects. There are three types of Processes: Defining, Realizing, and Implementing. In addition, physical Objects include Aircraft, System, Item, and Item component.

D. Functional Resonance Analysis Method

Functional Resonance Analysis Method (FRAM) [13] has been used to analyze complex functional resonances of socio-technical systems through functional networks. The FRAM function is defined by hexagonal nodes with six sides. These sides correspond to six aspects which are Input, Output, Time, Control, Resource, and Precondition. The output side of a function can be connected to the other five sides of other functions. FRAM provides useful means for safety analysis. Possible aspect relationships are <O, I>, <O, T>, <O, C>, <O, R>, and <O, P>. Here, <X, Y> is where X and Y are functional aspects.

The following three types of FRAM matrix representations have been proposed.

Lundberg and Woltjer [14] proposed a Resilience Analysis Matrix (RAM) to visualize functional dependencies between complex systems. RAM is a square matrix that shows the propagation relationship between functions. The size of RAM is the number of functions in FRAM. Element (i, j) of RAM indicates that some aspect of function i is propagated from the output of function j. The diagonal element (i, i) of RAM is the output of function x.

Patriarca et al. [15] proposed another square matrix composed of aspect combinations of FRAM functions. If there are n couplings in FRAM, RAM is defined as an $n \times n$ square matrix. The value of RAM (i,j) is 1 or 0.

Functional Aspect Resonance Matrix (FARM) is a non-square matrix that shows the propagation relationship between the output of a function and other aspects. The number of rows in FARM is the number of output sides of the function that are propagated to other functions in FRAM. The column size of FARM is the number of sides of a

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

function that are connected to the output sides of other functions. Element (i, j) of FARM indicates that some functional surface j is propagated from the output of function i. In general, the number of rows and columns in FARM is not equal, so there are no diagonal elements. The equivalence of the above three matrices has been shown by Yamamoto [16].

E. Defect detection by visual feature analysis

A blog post [17] from dataspa.ai notes that visual inspection is a critical process in the manufacturing industry, where minute surface defects (scratches, cracks, peeling, etc.) have a significant impact on product quality. However, manual inspection is subjective and unscalable. Therefore, utilizing generative AI and synthetic data to artificially generate defect images can address data shortages. The synthetic data improves model accuracy and reliability, while reducing implementation costs.

A research paper [18] by Singh et al. comprehensively reviews synthetic image generation technologies to improve the performance of deep learning models for surface defect detection in the manufacturing industry. Deep learning models require large amounts of training data, but collecting defect images on-site is difficult and costly. On the other hand, CG-based approaches using CAD models and renderings lack the feasibility of defect generation.

These papers highlight the common challenge of data shortages in detection in the manufacturing industry, and the promise of synthetic data generation using generative AI as a solution.

However, the defect detection in manufacturing that these two documents focus on deals with "physical and visual features" such as images and sensor data, and defines defects as abnormalities in shape, color, or pattern.

However, defects in production and business processes arise from abstract and logical structures such as business flows, decision-making, and communication. Because these cannot be expressed as images or simple numerical patterns, they cannot be addressed by methods that rely on image generation or visual feature learning.

III. DEFECT PREVENTION MATRIX ANALYSIS

A. Defect Prevention Diagram

A defect prevention diagram consists of business processes and flow relationships between business processes. In a business process, input, output, accepting conditions, resource conditions, judgment conditions, and exception conditions are clarified. Flow relationships include flows from output to input and flows from exception conditions to input, resource conditions, and judgment conditions.

The Input describes the trigger and information for starting an action. The Output describes the response and information as a result of the action. Accepting conditions describe the conditions for executing an action. Resource conditions describe the people, equipment, information, and activities required to output the action results. Judgment conditions describe the criteria for outputting the action results. Exception conditions describe the conditions under which output cannot be generated because the receiving conditions, resource conditions, and judgment conditions are not met.

Figure 1 shows the defect prevention diagram process element.

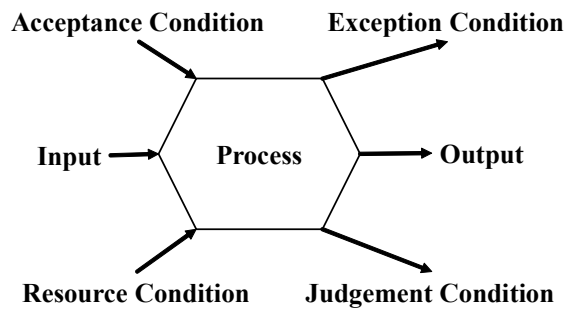


Figure 1. Defect prevention diagram process.

B. Defect Prevention Matrix

For the business process P that constitutes the defect prevention diagram D, the defect prevention matrix M can be defined as follows.

In Table I, S and T are either the receiving condition A, the resource condition R, or the judgment condition J. If S and T are omitted, they are taken to be the relationship to the input of the target process.

TABLE I. DEFECT PREVENTION MATRIX

	X	Y
X	Goal of X	X to S: Y Relationship
Y	Y to T: X Relationship	Goal of Y

The diagonal element M (X, X) describes the purpose of business process X. The off-diagonal element M (X, Y) describes the connection flow from business process X to either the input, receiving condition, resource condition, or judgment condition of Y.

The defect prevention matrix can be used to comprehensively check the connection flow between business processes that make up the defect prevention diagram. For example, the transitive closure of the defect prevention matrix can define a set of connection relationships for business processes. The set of connection relationships for X in Table I is $\Sigma_{k=1,n} (R_{xy} \cdot R_{yx})^k$. R_{xy} is the relationship from X to S: Y, and R_{yx} is the relationship from Y to T:X.

Similarly, the set of connection relationships for Y in Table I is $\Sigma_{k=1,n} (Ryx \cdot Rxy)^k$.

The defect prevention matrix is used to identify defects caused by the flow relationship among processes.

The scalability of the matrix approach depends on the complexity of the number of relations between processes. The approach can be adaptable for any business process relationships by using matrix representation.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

C. Process Review List

The process review checklist (PRC) is defined as issues of concern for six aspects, as shown in the following template in D.

D. Prompt template Defect Prevention Matrix Analysis

The prompt template is designed as follows:

The business process relationship matrix M can be defined as follows.

The diagonal element M (X, X) describes the purpose of business process X. The off-diagonal element M (X, Y) represents the connection flow from business process X to either the input, receiving condition, resource condition, or judgment condition of Y.

Develop M for the following < descriptions>.

<Body of description>

By using the following checklist, identify defects in the M processes.

Checklist:

The process review checklist is defined as issues of concern for the six aspects, as follows.

[<Process name>]

[Input] issues with input labels

[Accepting condition] issues with accepting arrow labels

[Resource condition] issues on resource arrow labels

[Judgement condition] issues with judgement arrow labels

[Output] issues with output arrow labels

[Exception condition] issues on exception arrow labels

IV. EXPERIMENTAL STUDY

A. Shinkansen Bogie Inspection System

The Shinkansen bogie inspection system is a system implemented by JR Central to inspect Shinkansen bogies in order to maintain safe Shinkansen operation. The main components of the Shinkansen bogie inspection system are (1) the Shinkansen bogie to be inspected, (2) the maintenance workers who check for abnormalities in the bogie, (3) the controller who issues orders for the bogie inspection, and (4) the supervisor who manages the entire train operation management process.

The above description of the Shinkansen bogie inspection system is provided to the prompt template of the Defect Prevention Matrix Analysis. Then the following response is generated.

To develop the **Defect Prevention Matrix (M)** for the **Shinkansen bogie inspection system**, Human first identify the key business processes based on the components AI listed:

- **Bogie Inspection** (performed on the Shinkansen bogie)
- **Maintenance Work** (performed by maintenance workers)
- **Inspection Control** (managed by the controller)
- **Operation Supervision** (managed by the supervisor)

TABLE II. DEFECT PREVENTION MATRIX M

M (X, Y)	Bogie Inspection (BI)	Maintenance Work (MW)	Inspection Control (IC)	Operation Supervision (OS)
Bogie Inspection (BI)	Purpose: To assess the physical condition of the Shinkansen bogie for abnormalities.	Provides inspection results as input for maintenance decisions.	Receives inspection orders from IC.	Supplies inspection data for operational oversight.
Maintenance Work (MW)	Uses inspection data to perform necessary repairs or adjustments.	Purpose: To correct any abnormalities found in the bogie.	Receives task assignments based on IC decisions.	Reports maintenance status for operational tracking.
Inspection Control (IC)	Issues inspection orders and schedules.	Sends instructions and resource allocations to MW.	Purpose: To manage and coordinate inspection activities.	Receives strategic directives and constraints from OS.
Operation Supervision (OS)	Uses inspection and maintenance data to oversee train safety and scheduling.	Evaluates maintenance outcomes to ensure operational readiness.	Provides high-level directives and judgment conditions to IC.	Purpose: To ensure safe and efficient train operation through oversight of all processes.

Generative AI defines the matrix **M** as shown in Table II, where:

- **Diagonal elements M (X, X)** describe the **purpose** of each process.
- **Off-diagonal elements of M (X, Y)** describe the **connection flow** from process X to process Y (e.g., input, receiving condition, resource condition, or judgment condition).

To identify **defects** in the business processes of the Shinkansen bogie inspection system using the **process review checklist**, generative AI evaluates each process across the six aspects. Issues detected for business processes by generative AI are shown in Tables III through VI, respectively. Each Table corresponding to business processes shows issues for six aspects.

TABLE III. BOGIE INSPECTION (BI) TABLE

Aspect	Issues
Input	Lack of standardized data formats from previous inspections. Incomplete historical maintenance records.
Accepting condition	No clear criteria for when a bogie is ready for inspection (e.g., post-trip, time-based). Manual confirmation may lead to delays or errors.
Resource Condition	Limited availability of inspection tools or sensors. Dependency on human inspectors without automation support.
Judgement condition	Subjective assessment by inspectors may vary. No AI or rule-based system to assist in anomaly detection.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Output	Inspection results may be inconsistently documented. Lack of integration with maintenance scheduling systems.
Exception condition	No protocol for handling unexpected bogie conditions (e.g., severe damage). Delays in reporting critical findings to supervisors.

TABLE IV. MAINTENANCE WORK CHECK

Aspect	Issues
Input	Inspection data may be unclear or lack actionable recommendations. Missing parts or repair history.
Accepting condition	No automated system to verify readiness for maintenance. Manual task assignment may cause bottlenecks.
Resource Condition	Shortage of skilled workers or spare parts. Equipment downtime.
Judgement condition	Decisions on repair priority may lack standardization. No predictive maintenance model in place.
Output	Maintenance logs may not be digitized or standardized. No feedback loop to inspection or control processes.
Exception condition	Emergency repairs may not follow standard procedures. Lack of escalation protocol for unresolved issues.

TABLE V. INSPECTION CONTROL CHECK

Aspect	Issues
Input	Incomplete data from operation supervision or previous inspections. Lack of real-time bogie status.
Accepting condition	No validation of inspection readiness. Manual scheduling is prone to errors.
Resource Condition	Limited access to real-time data or analytics. Over-reliance on manual coordination.
Judgement condition	No decision support system for prioritizing inspections. Lack of integration with train operation schedules.
Output	Orders may be delayed or miscommunicated. No audit trail for issued commands.
Exception condition	No contingency plan for inspection delays. Inadequate handling of conflicting inspection schedules.

TABLE VI. OPERATION SUPERVISION (OS) CHECK

Aspect	Issues
Input	Delayed or incomplete reports from inspection and maintenance. No real-time dashboard for system status.
Accepting condition	No automated alerts for critical issues. Manual review of reports may cause oversight.
Resource Condition	Limited access to integrated system data. Overload of supervisory responsibilities.
Judgement condition	Decisions may be reactive rather than predictive. No simulation tools for operational impact analysis.
Output	Strategic directives may lack clarity or timeliness. No feedback mechanism to lower-level processes.
Exception condition	No protocol for emergency train operation decisions. Lack of coordination during system-wide disruptions.

B. Insulin Pump Control System

Next, we applied and reviewed the method to an insulin pump control system [19].

Insulin pump requirements sentences are as follows;

R1) Insulin pump control software manages sensors implanted in the patient to measure blood parameters proportional to blood glucose levels.

R2) Blood parameters are sent to the pump controller.

R3) The controller calculates the sugar content and amount of insulin needed.

R4) The controller sends a signal to a small pump to administer insulin via a needle implanted in the patient.

R5) Insulin pump delivers 1 unit of insulin in response to 1 unit pulse from controller.

By using the Defect Prevention Matrix Analysis template, Generative AI breaks down the requirements into identifiable business processes:

BP1) Measure Blood Parameters (from R1)

BP2) Send Blood Parameters (from R2)

BP3) Calculate Insulin Needs (from R3)

BP4) Send Insulin Administration Signal (from R4)

BP5) Administer Insulin (from R5)

The key changes extracted are as follows.

a) Explicit Inputs and Resources: More precise descriptions of what each process receives and what it uses to operate.

b) Clearer Judgment Conditions: Adding checks or criteria that must be met for a process to proceed or complete successfully.

c) Crucial Exception Conditions: Highlighting potential failure points that require handling, essential for medical devices.

Table VII shows the elements of the Defect Prevention Matrix for the Insulin pump requirements. The matrix is shown in the form of purpose, conditions and flow columns. The matrix is shown in the form of purpose, conditions and flow columns. The purposes and conditions of diagonal elements are described in the raw. The off-diagonal elements are represented in flow columns. The form is selected by the sake of space efficiency.

TABLE VII. DEFECT PREVENTION MATRIX OF INSULIN PUMP REQUIREMENTS

	PURPOSE	CONDITIONS	FLOW
BP1	Continuously measures blood parameters using implanted sensors.	**Resources:** Functional implanted sensor. **Exceptions:** Sensor malfunction, disconnection, erroneous readings, inability to obtain a reading.	**Flow to BP2:** Measured raw blood parameter data sent to the pump controller.
BP2	Transmits the measured raw blood parameter data from the sensor to the pump controller	**Resources:** Operational communication module/channel (sensor to controller) **Exceptions:** Communication failure (signal loss, interference), data corruption during transmission.	**Flow to BP3:** Valid and complete blood parameter data successfully received from the sensor.
BP3	Processes received blood parameter data to calculate current glucose levels and determine the precise amount of insulin required.	**Resources:** Pump controller's processing unit, pre-programmed calculation algorithm. **Judgment Condition:** Calculated glucose level is within processing range; Calculated insulin amount is within pre-defined safe dosage limits	**Flow to BP4:** Insulin administration signal containing the calculated dosage amount, ready for transmission to the pump.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

		Exceptions: Calculation error (e.g., invalid input, algorithm failure), calculated insulin amount is dangerously high or low, inability to determine a safe dose	
BP4	Transmits the calculated insulin administration signal from the controller to the insulin pump mechanism.	**Resources:** Operational communication module/channel (controller to pump). **Exceptions:** Signal transmission failure to the pump, signal corruption, non-receipt by the pump	**Flow to BP5:** Valid insulin administration signal/pulse successfully received from the controller.
BP5	Mechanically delivers the precise amount of insulin into the patient's body via the implanted needle based on the received signal.	**Resources:** Functional small pump mechanism, implanted needle/infusion set, **sufficient insulin in reservoir.** **Judgment Condition:** Sufficient insulin available in reservoir; Pump mechanism is operational (no blockages, sufficient battery); Needle/infusion set is properly functioning (not blocked/dislodged). **Exceptions:** Low or empty insulin reservoir, occlusion/blockage of needle/tubing, dislodged or bent needle, pump mechanical failure, low battery, detected over-infusion or under-infusion, patient adverse reaction (requiring immediate pump stop/alarm).	

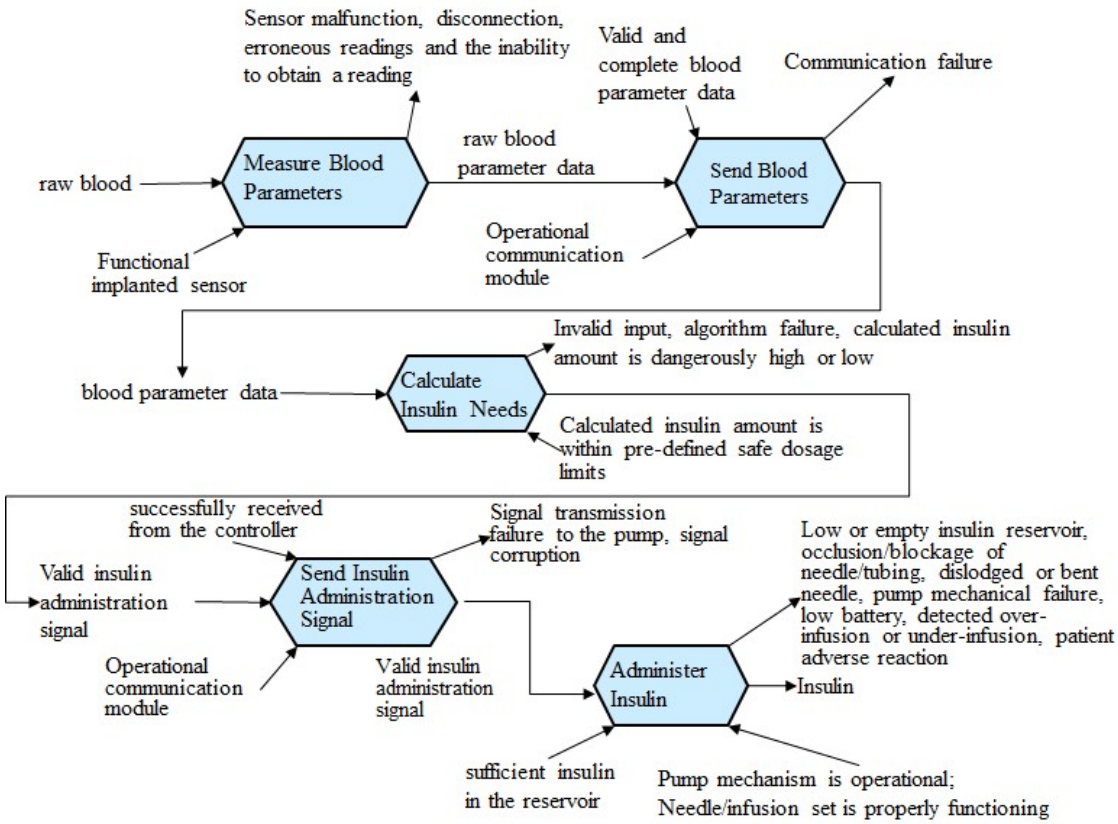


Figure 2. Defect Prevention Diagram of Insulin Pump Control System

Figure 2 shows the Defect Prevention Diagram based on the result of the Defect Prevention Matrix Analysis for the Insulin Pump Control System requirements.

C. PDCA

The following PDCA process description is then applied to the Prompt template DPMA.

Process Description: PDCA

Plan: Set goals and plan ways to achieve them.

Do: Take action based on the plan.

Check: Check the results and assess whether they went according to plan.

Act: Take improvement measures based on the evaluation and incorporate them into the next plan.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Generative AI developed the Defect Prevention Matrix as shown in Table VIII. Then the check results for the plan, do, check, and act process are shown in Tables IX to XI.

The DPM of PDCA clarifies the inter-relationship among the four processes of PDCA. The Table indicates that non-adjacent processes also have dependencies in the PDCA cycle. For example, there are three backward loops from the Do, Check, and Act processes to the Plan process. In addition, the PDCA cycle usually only shows the loop from Act to Plan, but the Table shows that Plan also has a relationship with Check and Act.

TABLE VIII. DEFECT PREVENTION MATRIX OF PDCA

M(X, Y)	Plan (P)	Do (D)	Check (C)	Act (A)
Plan (P)	Purpose: Set goals and plan ways to achieve them.	Provides input and resource conditions for Do.	Sets judgment criteria for Check.	Provides baseline for improvement in Act.
Do (D)	Uses plan to execute actions.	Purpose: Take action based on the plan.	Supplies results for evaluation in Check.	Provides operational data for Act.
Check (C)	Uses Do's output to evaluate performance.	Assesses execution against Plan.	Purpose: Check results and assess alignment with plan.	Provides evaluation results for Act.
Act (A)	Uses Check's evaluation to improve future plans.	Refines execution strategy for Do.	Adjusts judgment criteria for Check.	Purpose: Take improvement measures and feed them into the next Plan cycle.

Tables IX to XII show the checklists of PDCA processes. The checklists clarify the successful conditions of the PDCA cycle. If these successful conditions are not clear, the PDCA cycle will result in unnecessary rework. However, these success criteria seem not to have been given much importance when the PDCA cycle was first introduced.

Defect prevention analysis clarifies the success conditions for the PDCA process, promoting an efficient PDCA cycle. It also enables quick responses to exceptions that occur in each PDCA process, reducing the number of iterations of the outer loop of PDCA.

Therefore, by integrating defect prevention analysis with the PDCA cycle, business improvement through PDCA can be made more efficient. Defect prevention matrix analysis can clearly define the relationships between the processes that make up the PDCA cycle, which may allow for the prediction of PDCA cycle iterations in advance.

TABLE IX. CHECKLIST OF THE PLAN PROCESS

Aspect	Issues
Input	Goals may be vague or not measurable. Lack of historical data or context.
Accepting condition	No formal criteria for initiating planning. Planning may begin without stakeholder alignment.
Resource Condition	Insufficient tools or data for effective planning. Lack of cross-functional input.
Judgement condition	No clear success metrics defined. Risk assessment may be missing.
Output	Plans may be too abstract or not actionable. No version control or documentation standards.
Exception condition	No fallback strategy if planning fails. Unclear escalation path for unresolved planning conflicts.

TABLE X. CHECKLIST OF DO PROCESS

Aspect	Issues
Input	Plans may be misinterpreted or incomplete. Lack of clarity in task assignments.
Accepting condition	No readiness check before execution. Missing confirmation of resource availability.
Resource Condition	Tools or personnel may be underprepared. No contingency resources.
Judgement condition	Execution decisions may lack flexibility. No real-time monitoring or feedback.
Output	Results may not be recorded systematically. No linkage to performance indicators.
Exception condition	No protocol for handling execution failures. Delays may not be communicated upstream.

TABLE XI. CHECKLIST OF CHECK PROCESS

Aspect	Issues
Input	Data from Do may be incomplete or inconsistent. No baseline for comparison
Accepting condition	No criteria for when to initiate evaluation. Evaluation may be skipped under time pressure.
Resource Condition	Lack of analytical tools or expertise. No access to historical benchmarks.
Judgement condition	Evaluation may be biased or subjective. No structured review process.
Output	Findings may not be actionable. No feedback loop to Plan or Act.
Exception condition	No handling of conflicting evaluation results. Errors in data may go unnoticed.

TABLE XII. CHECKLIST OF ACT PROCESS

Aspect	Issues
Input	Evaluation results may be unclear or delayed. No prioritization of improvement actions.
Accepting condition	No criteria for implementing changes. Resistance to change may block action.
Resource Condition	Lack of authority or resources to implement changes. No support for training or adaptation.
Judgement condition	Decisions may be reactive rather than strategic. No long-term impact analysis.
Output	Improvements may not be documented or tracked. No integration into future planning.
Exception condition	No plan for failed improvement attempts. Feedback may not reach the right stakeholders.

V. DISCUSSION

A. Novelty

In this paper, we have proposed a procedure for creating the defect prevention analysis matrix and a review method, as well as an automatic execution procedure using generative AI.

1) Structuring Business Process Knowledge

In a defect prevention diagram, business process knowledge can be organized hierarchically using L1: business process knowledge, L2: business flow-related knowledge, L3: business process action conditional knowledge, and L4: business action condition execution knowledge. Here, L1, L2, and L3 can be described in a defect prevention diagram. However, for L4, the described

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

conditions must be evaluated when the actual business process is executed.

In the business process, knowledge of a defect prevention diagram, L1 can grasp the overall picture of the business process by identifying the necessary actions that make up the business process. Business flow-related knowledge L2 can recognize the dependencies between business processes. Business process action condition knowledge L3 can recognize the necessary conditions to execute a business. The difference between L3 and L4 is the difference between knowing the conditions and being able to appropriately confirm and evaluate those conditions. Condition evaluation knowledge L4 should be specified so that the evaluation results do not vary depending on the individual for the same conditions.

In the defect prevention diagram, this type of business process knowledge classification is used to organize business knowledge that has traditionally been thought to vary between individuals, making it possible to clarify where these variations in knowledge occur.

The defect prevention diagram was originally proposed as the name of the business process completeness diagram [20].

2) Automating defect review with generative AI

As shown in the application example, the proposed Prompt template easily automates the review of business processes using the defect prevention matrix, demonstrating the applicability of the proposed method.

In the defect prevention diagram shown in Figure 3, created by humans [16], the details of the exceptions were insufficient. In contrast, Tables III to VI extracted by the generative AI show that specific exception candidates were elicited.

B. Applicability

In this paper, we confirmed the applicability of the proposed method by applying it to the train operation monitoring process, the PDCA cycle, and the insulin pump control system. Because these cases are important business processes for safety operation monitoring, the proposed method may apply to a wider range of safety-critical processes.

Furthermore, as we can see from Figure 2, the defect prevention diagram of the insulin pump control system only has linear relationships between the preceding and following processes, whereas the process relationships of the Shinkansen bogie inspection system and the PDCA cycle have comprehensive interactive relationships between the processes.

C. Comparison with RCA and FRAM

In Root Cause Analysis (RCA), when a defect is detected in a system, the cause of the defect is identified. Once the cause is identified, measures are devised to prevent the defect from occurring in the system. In contrast, in defect prevention analysis, which is the premise of the defect prevention diagram, the success conditions and exceptions of the system are first defined. Next, measures to deal with exceptions are devised in the system. Defects that occur during the operation of the system are identified, and the planned measures are implemented.

FRAM and the Defect Prevention Matrix Analysis (DPMA) have common aspects as input, output, and resources. FRAM has time, precondition, and control aspects that are not in DPMA. DPMA also has acceptance, exception, and judgment condition aspects, which are not in FRAM. The output of FRAM is restricted to output aspects. Therefore, the meaning of output in FRAM may be unclear as it is difficult to discriminate exceptional output from normal output by aspects.

Although there are differences between FRAM and DPMA, it is unclear whether they have the same expressive power. As FRAM can be applied to analyze the resonance relationship between processes, the completeness of

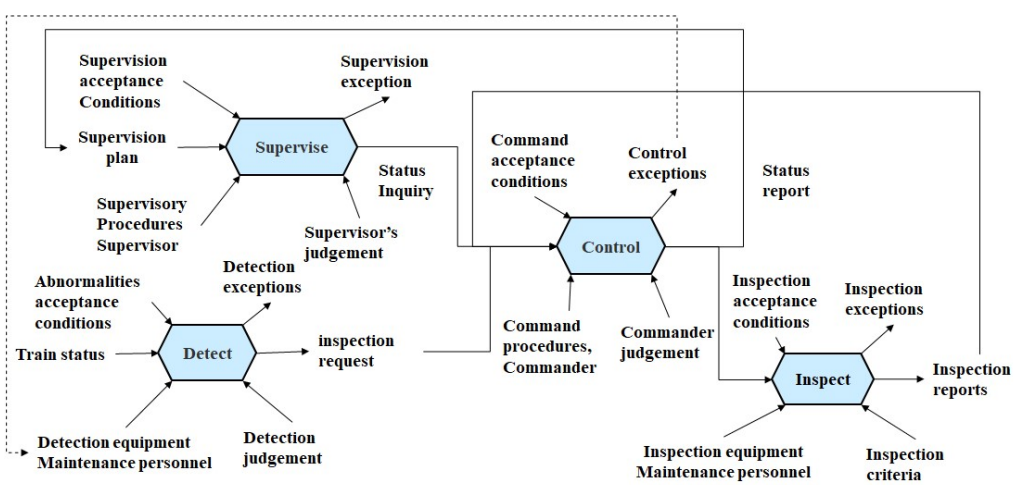


Figure 3. Defect Prevention Diagram of Shinkansen Bogie Inspection System

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

business processes may also be possible to analyzed by FRAM. Sujan and Felici [21] combine Failure Mode and FRAM. This implies a new method possibility that integrates the analysis method using DPMA with Failure mode analysis. FRAM attempts to model the diversity and interactions of a system in detail, but if the model is too detailed, it becomes overly complex, making analysis difficult and impractical insights difficult to obtain. It's difficult to determine how much diversity to describe and at what granularity to define functions, and excessive modeling can increase analysis time and cause essential insights to be overlooked.

Even if the mechanism of variation can be identified, specific countermeasures may not be obvious. Further consideration is required to identify the most effective interventions among the complex interactions.

FRAM is not simply a template-based approach; it requires a skilled facilitator who can deeply understand and facilitate discussion about the functions that make up the system, their interactions, and the sources of diversity. Attempting to apply FRAM without experience can result in a superficial analysis or overlooking important aspects, resulting in ineffective results.

The advantages of DPMA over FARM are as follows:

1) DPMA can derive success conditions based on the business process definition. FARM requires identifying functions from the business process, which not only leaves the granularity of the function definition indefinite, but also makes the optimal granularity dependent on personal judgment. Furthermore, FRAM lacks clear standards for extracting success conditions, making consensus-building among stakeholders essential.

2) FRAM explores success and failure conditions based on functional resonance, assuming that the execution conditions of functions are unclear, making it difficult to ensure the integrity of business processes. DPMA clearly defines process integrity conditions, allowing deviations from those conditions to be detected and addressed as exceptions. Without defining normal conditions, it is impossible to guarantee the safety of business processes.

The formal comparison between FRAM and DPMA is an interesting future research theme.

Table XIII shows the comparison of RCA and FRAM to DPMA.

TABLE XIII. COMPARISON WITH RCA AND FRAM

	RCA	FRAM	DPMA
Overview	Investigate the cause of the problem	Success-based integration	Comprehensive Defect Prevention
Direction	Backward analysis	Upstream/downstream analysis	Forward/ Backward Analysis
Implementation Period	Post-incident	Pre- and post-event analysis	Proactive
Countermeasures	Prevent recurrence	Prevention of recurrence	Prevention
Focus Points	Occurring problem	Functional fluctuation	Success Conditions, Exceptions, and Responses
Methods	Cause-and-effect analysis	Functional resonance analysis	Exception Propagation Analysis
Challenges	Expand to other problems	Identification and comprehensiveness of aspects and resonance relationships	Overall Optimization and Residual Risk

D. Exception propagation

We need to clarify how to respond to process exceptions. In the case study, we identified exception detection but not propagation. We can define a template to perform the exception propagation procedure as shown below. The template that performs the exception propagation procedure is as follows.

Prompt template: Exception propagation.

Identify the exception propagation relationships between processes for <the Processes>.
The exception propagation procedure is as follows:
The exceptional propagation relationships between processes are as follows:
Propagation to the receiving condition of the preceding process. Propagation to the decision condition of the preceding process. Propagation to the resource condition of the current process. Propagation to the receiving condition of the subsequent process. Propagation to the decision condition of the subsequent process

The template is applied to the PDCA process. The results of applying for the PDCA process are shown below.

TABLE XIV. Exception Propagation of PDCA cycle

Process	To	Exception	Impact
Plan	Resource Condition of Current Process	Lack of data, unclear goals, or unavailable planning tools	Planning cannot proceed effectively
Plan	Receiving Condition of Subsequent Process (Do)	Ambiguous success criteria	Execution may be delayed or misaligned
Plan	Decision Condition of Subsequent Process (Do)	Ambiguous success criteria.	Execution decisions may lack clarity or direction.
Do	Receiving Condition of Preceding Process (Plan)	Execution reveals flaws in the plan.	May require re-planning or plan revision.
Do	Decision Condition of Preceding Process (Plan)	Execution deviates from expected outcomes.	Planning assumptions may be invalidated.
Do	Resource Condition of Current Process	Lack of tools, personnel, or time.	Execution is compromised or halted.
Do	Receiving Condition of Subsequent Process (Check)	Poor documentation or missing data.	Evaluation becomes difficult or inaccurate.
Do	Decision Condition of Subsequent Process (Check)	Inconsistent results.	Evaluation may be inconclusive or misleading.
Check	Receiving Condition of Preceding Process (Do)	Evaluation identifies critical execution errors.	May require re-execution or correction.
Check	Receiving Condition of Preceding Process (Do)	Results contradict expected outcomes.	Execution decisions may be questioned.
Check	Resource Condition of Current Process	Lack of analytical tools or expertise.	Evaluation is incomplete or biased.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Check	Receiving Condition of Subsequent Process (Act)	Evaluation results are unclear or delayed.	Improvement actions may be postponed.
Check	Receiving Condition of Subsequent Process (Act)	Conflicting evaluation findings.	Decisions on improvements may be uncertain.
Act	Receiving Condition of Preceding Process (Check)	Improvement actions contradict evaluation.	May require re-evaluation.
Act	Decision Condition of Preceding Process (Check)	Changes invalidate previous assessments.	Evaluation criteria may need revision.
Act	Resource Condition of Current Process	Lack of authority, tools, or support for change.	Improvements are not implemented.
Act	Receiving Condition of Subsequent Process (Plan)	Feedback not incorporated into new plans.	Planning cycle does not evolve.
Act	Decision Condition of Subsequent Process (Plan)	Lessons learned not reflected in planning criteria.	Future plans repeat past mistakes.

The exception propagation for the Shinkansen bogie inspection system created by humans in Figure 3 does not specify exceptions and only extracts the exception propagation from the Control process to the Resource condition of the Defect process. In contrast, as shown in Tables III to VI, generative AI comprehensively extracts exceptions. Furthermore, in the exception propagation for the PDCA cycle, it extracts exception propagation not only to preceding and succeeding processes but also to the process itself. Therefore, exception propagation analysis by generative AI is more comprehensive than exception propagation created by humans.

Conversely, extracting comprehensive exception propagation has the problem of complicating the defect prevention diagram. Since it is not known whether all exception propagations will be necessary, the importance of exception propagation must be evaluated. In the future, a mechanism for evaluating the priority of exception propagation will need to be devised.

Another option is to stop all business processes when an exception occurs in order to respond to the exception. In terms of business process continuity, it is necessary to compare the method of continuing the execution of partial processes using exception propagation with the method of stopping all processes and responding to the exception. In the future, it will be necessary to consider criteria for evaluating the effectiveness of both methods.

E. Sufficiency of Six Process Aspects

The purpose of this paper is not to analyze complex system interactions, but to analyze defects in business processes. We propose extending the traditional self-process-containment method with exceptions and automating the procedure using generative AI. We do not claim that defects can be expressed using only six aspects. However, actual business processes can be expressed using the proposed five aspects and their exceptional aspects. Toyota has refined its manufacturing process to produce high-quality automobiles using a self-process-containment process. This demonstrates the practical sufficiency of the six aspects. Anyone with experience of designing actual business processes should be able to understand the sufficiency of the six aspects. Furthermore, our application to actual accident cases demonstrates that this is a practical defect analysis method.

The reasons why PRC based on the six aspects is not incomplete are as follows: First, it is always true whether the conditions of the five aspects are met. Second, an exceptional condition occurs when the five aspects are not met. Therefore, PRC based on the six aspects is sufficient.

F. Comparison with Learning Methods

The purpose of this paper is not to compare machine learning methods, but to demonstrate the possibility of automating the newly proposed business process defect analysis method using generative AI. Therefore, alternative methods for defect detection (such as supervised/unsupervised machine learning or GNNs) cannot replace the Defect Prevention Matrix. Generative AI is used as a means to automate defect analysis of process design in production and business processes. Copilot is used as the generative AI in this example. Similar results can also be obtained with ChatGPT and Gemini.

Comparison with business process defect detection methods is also made with methods such as FRAM and FMEA.

Furthermore, machine learning approaches require training data, which incurs unnecessary costs for use in business settings. Machine learning has strengths in detecting "visual defects" and "physical anomalies," but the same techniques cannot be applied to business process defects because they are non-visual, context-dependent, and dynamically defined.

In contrast, generative AI can execute procedures, eliminating redundant training costs. As described above, the purpose of this paper is to demonstrate the feasibility of using generative AI to automate defect analysis of maintenance and operation processes. Comparing the capabilities of generative AI and machine learning methods is beyond the scope of this paper and will be addressed in the future.

G. Limitation

In this paper, we have proposed a method for reviewing defect prevention diagrams. However, we have only applied it to a few cases. In the future, we need to quantitatively evaluate the effectiveness of the method by applying it to many cases.

However, because the number of exceptions detected by generative AI and their propagation will be large, we plan to continue studying how to select and discard them based on the importance of the detection results. We also need to clarify the criteria for determining whether to stop a business process when an exception is detected or to partially execute it based on the exception propagation response. Furthermore, since this paper only conducted a desk-based trial evaluation, it is necessary to apply it to more case studies and actual business processes for evaluation.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

VI. CONCLUSION

This paper proposes a review method using a defect prevention diagram, a process relationship table, and a checklist. The business process review checklist can verify the completeness of each of the six aspects of the processes that make up a defect prevention diagram. In particular, it can detect inconsistencies between multiple inputs and outputs. Furthermore, the process relationship matrix can analyze the comprehensive dependencies between the business processes that make up a defect prevention diagram.

By defining transition relationships based on the elements of the business relationship matrix M, it is possible to track influence relationships iteratively. In other words, M can be used to define a linguistic expression L for a defect prevention diagram. It is believed that this L can be used to formulate the equivalence of defect prevention diagrams, thereby minimizing the size of the defect prevention diagram.

The defect prevention diagram can complement the response to exceptions in business processes, making it possible to define business processes that can handle defects as exceptions.

This paper also clarifies prompt templates for a generative AI to automate the proposed method. Furthermore, the effectiveness of the proposed method was confirmed by applying it to a Shinkansen bogie inspection system, an insulin pump control system, and a PDCA cycle.

In this paper, we formulated the completeness of defect prevention diagrams in terms of their ability to respond to exceptions. We also demonstrated that comprehensive verification is possible through the automation of exception detection and exception propagation using generative AI.

However, because the number of exceptions detected by generative AI and their propagation will be large, we plan to continue studying how to select and discard them based on the importance of the detection results. We also need to clarify the criteria for determining whether to stop a business process when an exception is detected or to partially execute it based on the exception propagation response. Furthermore, since this paper only conducted a desk-based trial evaluation, it is necessary to apply it to more case studies and actual business processes for evaluation.

REFERENCES

[1] S. Yamamoto, "Defect Prevention Review by Process Relationship Matrix" The Seventeenth International Conference on Information, Process, and Knowledge Management (eKNOW 2025) IARIA, 2025, pp. 37-40.

[2] S. Sasaki, "Self-process completion - Quality is built in the process," JSQC selection, Japan Society for Quality Control, 2014. (in Japanese)

[3] J. Salvadorinhoa and L. Teixeira, "Organizational knowledge in the I4.0 using BPMN: a case study," CENTERIS 2021, Procedia Computer Science 181, 2021, pp. 981–988.

[4] A. Ore, O. Kuznecova, and A. Jegorova, "Self-managed Organization: A Role of Business Process Management," 2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS), doi: 10.1109/ITMS52826.2021.9615260

[5] P. Faraboschi, E. Frachetenberg, P. Laplante, D. Milojicic, and R. Saracco, "Digital Transformation: Lights and Shadows," IEEE Computer, 2023, pp. 123-130.

[6] S. Yamamoto, "A Strategic Map for Digital Transformation," KES 2020, Procedia Computer Science, vol. 176, 2020, pp. 1374-1381.

[7] S. Yamamoto, "A Comparative Analysis of Business Model Notations," Journal of Business Theory and Practice, ISSN 2372-9759 (Print) ISSN 2329-2644 (Online), vol. 7, no. 3, 2019.

[8] D. Leonard and W. Swap, "Deep Smarts: How to Cultivate and Transfer Enduring Business Wisdom," Harvard Business Review Press, 2005.

[9] Z. Nakamura, "How to conceive a simple job—Succeed in Entity Process Analysis," Nikkan Kogyo Shimbun, 2003 (in Japanese)

[10] S. Yamamoto and H. Fujimoto, "PKC-- Production Knowledge Chart for Knowledge Transfer," KES2023, Procedia Computer Science 225, 2023, pp. 414-423.

[11] D. Dori, "Model-Based Systems Engineering with OPM and SysML," Springer, 2016.

[12] B. Cameron, E. Crawly, and D. Selva, "System Architecture," PEARSON, 2015.

[13] E. Hollnagel, "FRAM - the Functional Resonance Analysis Method: Modelling Complex Socio-Technical Systems." CRC Press, 2012.

[14] J. Lundberg, R. Woltjer, "The Resilience Analysis Matrix (RAM): Visualizing functional dependencies in complex socio-technical systems," In: Proceedings of the 5th Resilience Engineering Association Symposium, 2013.

[15] R. Patriarca, G. Pinto, G. Di Gravio, F. Costantino, "FRAM for Systemic Accident Analysis: A Matrix Representation of Functional Resonance," International Journal of Reliability Quality and Safety Engineering, vol. 25, no. 1, 2018, doi: 10.1142/S0218539318500018

[16] S. Yamamoto, "Comparative Study on Functional Resonance Matrices," Knowledge-Based Software Engineering: 2022: Proceedings of the 14th International Joint Conference on Knowledge-Based Software Engineering (JCKBSE 2022), M. Virvou, T. Saruwatari, L. Jain *Editors,* Learning and Analytics in Intelligent Systems 30, Springer, 2023, pp. 169-180, doi:org/10.1007/978-3-031-17583-1

[17] dataspan.ai blog, Automated Defect Detection: How GenAI and Synthetic Data are Transforming Visual Inspection in Manufacturing, July 2025, https://www.dataspan.ai/blog/automated-defect-detection-how-genai-and-synthetic-data-are-transforming-visual-inspection-in-manufacturing.

[18] A. R. Singh and others, A comprehensive survey of image synthesis approaches for Deep Learning-based surface defect detection in manufacturing, Computers in Industry, ScienceDirect Elsivier, Computers in Industry, Volume 173, ISSN 0166-3615, https://doi.org/10.1016/j.compind.2025.104360.2025.

[19] I. Sommerville, Software Engineering Tenth Edition, Addison-Wesley, 2015.

[20] S. Yamamoto, "Business Process Completeness" The Sixteenth International Conference on Information, Process, and Knowledge Management (eKNOW 2024) IARIA, 2024, pp. 24-28.

[21] M. Sujan and M. Felici, "Combining Failure Mode and Functional Resonance Analyses in Healthcare Settings," SAFECOMP, LNCS 7612, 2012, pp. 364-375.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

From Bias to Balance: Fairness-Aware Paper Recommendation for Equitable Peer Review

Uttamasha Anjally Oyshi, Susan Gauch
Department of Electrical Engineering & Computer Science
University of Arkansas
Fayetteville, USA
e-mails: {uoyshi, sgauch}@uark.edu

***Abstract*—Despite frequent double-blind review, systemic biases related to author demographics still disadvantage underrepresented groups. We start from a simple hypothesis: if a post-review recommender is trained with an explicit fairness regularizer, it should increase inclusion *without* degrading quality. To test this, we introduce *Fair-PaperRec*, a Multi-Layer Perceptron (MLP) with a differentiable fairness loss over intersectional attributes (e.g., race, country) that re-ranks papers after double-blind review. We first probe the hypothesis on *synthetic* datasets spanning high, moderate, and near-fair biases. Across multiple randomized runs, these controlled studies map where increasing the fairness weight strengthens macro/micro diversity while keeping utility approximately stable, demonstrating robustness and adaptability under varying disparity levels. We then carry the hypothesis into the *original* setting—conference data from ACM Special Interest Group on Computer-Human Interaction (SIGCHI), Designing Interactive Systems (DIS), and Intelligent User Interfaces (IUI). In this real-world scenario, an appropriately tuned configuration of Fair-PaperRec achieves *up to* a 42.03% increase in underrepresented-group participation with at most a 3.16% change in overall utility relative to the historical selection. Taken together, the synthetic-to-original progression shows that fairness regularization can act as both an equity mechanism and a mild quality regularizer, especially in highly biased regimes. By first analyzing the behavior of the fairness parameters under controlled conditions and then validating them on real submissions, Fair-PaperRec offers a practical, equity-focused framework for post-review paper selection that preserves—and in some settings can even enhance—measured scholarly quality.**

Keywords-Fairness-aware recommendation; Paper selection; Demographic bias mitigation

I. Introduction

This journal version extends our earlier conference paper [1], offering a deeper and more comprehensive analysis of fairness-aware learning for mitigating post-review bias in academic peer review. Double-blind review often does not eradicate systemic biases linked to authors' demographics, reputations, or institutional affiliations, despite attempts to ensure impartiality [2], [3], [4], [5]. Recent data indicate that even the most stringent anonymization techniques can be undermined by analyzing writing style or cross-referencing previous articles [6], [7]. This tendency can sustain biases against particular groups, including women, racial minorities, and researchers from underrepresented areas [4], [8], [9], [10]. Simultaneously, there is a growing dependence on recommendation algorithms to optimize processes such as paper selection, grant distribution, and significant publication identification [11], [12], [13]. While these systems can accelerate decision-making, they also pose a danger of perpetuating biases present in the training data, particularly if they focus only on predictive accuracy [14], [15], [16]. Therefore, it is imperative to devise novel methodologies that explicitly include demographic justice, preventing the perpetuation of historical inequalities.

We start from a simple, testable hypothesis on synthetic data: increasing the strength of fairness regularization (λ) should improve inclusion with limited impact on utility, and there should exist a sweet-spot range that depends on the underlying bias level. We map this hypothesis across controlled regimes (fair, moderate, high bias) for multiple protected attributes (race, country), then carry the learned settings into *original* conference data (ACM SIGCHI, DIS, IUI) to validate external relevance. This two-stage progression, from controlled what-if to real-world does-it-hold, reveals when fairness regularization uncovers under-selected high-quality work and when larger λ values begin to over-correct a system that is already close to balanced.

In this paper, we introduce *Fair-PaperRec*, a fairness-aware recommendation framework specifically designed to mitigate *post-review* bias through a differentiable fairness loss integrated with a prediction objective. Unlike heuristic approaches that often handle single-attribute constraints or overlook intersectionality, our approach:

- Surpasses single-attribute approaches by incorporating multiple demographic attributes (e.g., race, country) and constructing profiles that capture underlying biases.

This is an extended version of: U. A. Oyshi and S. Gauch, "Fair Learning for Bias Mitigation and Quality Optimization in Paper Recommendation," 17th Annual Conference on Information, Process, and Knowledge Management (eKNOW 2025), Nice, France, May 18–22, 2025.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

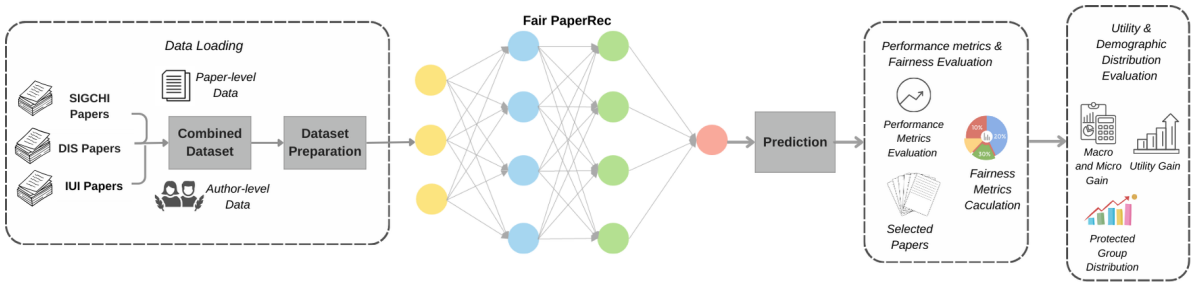


Figure 1. Overview of the Fair-PaperRec Architecture.

- After double-blind review, a specialized fairness penalty is implemented to address demographic disparities, correcting latent biases without replacing existing processes.
- The method explicitly targets demographic parity while monitoring utility, achieving equitable representation without compromising academic rigor.

Our results demonstrate improved representation in the participation of underrepresented groups, as well as an enhancement in overall paper quality. Notably, these findings reveal that enhanced inclusivity need not diminish academic rigor; a fairness-driven approach can yield greater demographic parity while simultaneously preserving, and at times even *enhancing*, the quality of accepted papers.

Contributions: This paper makes three contributions:

1) *A post-review, multi-attribute fairness framework* that integrates a differentiable fairness loss with prediction to address intersectional disparities while preserving quality.
2) *A two-stage evaluation methodology* that (i) charts fairness-utility trade-offs on synthetic bias regimes to identify λ sweet spots, then (ii) validates transfer to original conference data (SIGCHI, DIS, IUI).
3) *Actionable guidance* for choosing λ and attributing weights based on disparity levels, showing when fairness delivers win-win improvements and when smaller regularization suffices.

By mitigating biases in paper selection, our strategy promotes a richer academic discourse and amplifies the representation of marginalized communities, thereby paving the way toward more equitable, high-quality conferences. The paper is organized as follows: Section II reviews related work; Section III presents the methodology; Section IV details the experimental setup and metrics; Section V reports synthetic and original results with analysis; and Section VI concludes.

II. Related Work

We review three strands of literature to motivate our two-stage story, starting from a simple hypothesis on synthetic data and ending in a real-world conclusion on original conference data. First, we examine where and why double-blind review can fail, emphasizing structural incentives beyond anonymity. Second, we connect these pressures to fairness in recommendation and ranking. Third, we cover post-review mitigation and neural approaches most relevant to our setting.

A. Double-Blind Review, Structural Incentives, and Bias in Academic Selection

Although double-blind review conceals identities [2], [3], [4], it often fails to eliminate disparities by gender, race, or geography [10], [17]. Re-identification via stylometry or linkage to prior work can undermine anonymity [6], while prestige effects still favor well-known institutions [10]. Consequently, underrepresented groups remain disadvantaged [18], and acceptance gaps persist [8], [9].

Beyond anonymity, *structural incentives* can shape outcomes. Conflicts of interest (COIs) and funding ties have become more visible as industry participation in ML/NLP expands. Recent large-scale audits report the prevalence and statistical correlates of industry involvement across venues and years [19], [20], with complementary bibliometric evidence that industry and academia contribute differently across problem types, data assets, and deployment pathways [21], [22], [23]. Co-authorship and collaboration networks also influence visibility and impact [24], which can indirectly translate into acceptance advantages. Field norms further matter: in computing, the centrality of *conferences* (deadlines, rapid iteration) versus *journals* can shape incentives and review practices [25], [26]. Together, these factors motivate *post-review* fairness controls that are sensitive to venue norms and network effects, not just anonymization fidelity.

Because peer selection often conflates contribution and credit, authorship order practices vary across subfields and team sizes [27], [28]. Likewise, h-index is field- and seniority-dependent, with wide normative ranges and known limitations as a sole quality proxy [29], [30]. These observations motivate our decision to (i) treat author-level "utility" as a *weighted* h-index by

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

career stage and (ii) evaluate fairness at *paper* and *author* levels (macro/micro).

B. Fairness in Recommendation and Ranking

When optimizing solely for accuracy, recommenders can amplify historical bias [12], [31]. Multi-objective [14], adversarial [15], and re-ranking approaches [16] offer alternatives but frequently target a *single* protected attribute or generic user–item contexts, leaving intersectional and post-review constraints underexplored. In scholarly selection, *provider fairness* aligns with *author fairness*; the focus is on equitable exposure and acceptance for authors, not only on predictive accuracy [32]. Contemporary venue trends (e.g., topic shifts and data/benchmark access highlighted at CVPR 2024) underscore how exposure and resources interact with selection signals [33], reinforcing the need to treat fairness and quality jointly.

In recommender-systems literature, *provider fairness* refers to ensuring equitable exposure or selection opportunities for content creators. In the academic peer-review context, the "providers" are the authors themselves; thus, provider fairness corresponds directly to *author fairness*, aiming to equalize acceptance probabilities across demographic groups. This distinction clarifies that our work focuses on fairness for authors (providers), rather than fairness for users of the recommendation system.

Our synthetic-to-original narrative builds on this literature by (i) explicitly modeling *multiple* protected attributes (race, country), (ii) mapping fairness-utility trade-offs across tunable bias regimes to locate λ "sweet spots," and (iii) transferring those settings to real conference data where industry/academia dynamics, network effects, and venue norms co-exist [21], [22], [23], [25], [26].

C. Post-Review Bias Mitigation and Neural Approaches

Heuristic post-review balancing can improve representation [32] but risks local optima and often omits multi-attribute fairness. Neural approaches such as *DeepFair* [12] and *Neural Fair Collaborative Filtering* [34] show that fairness and accuracy can align in commercial domains, yet peer review imposes hard quotas, reputational stakes, and limited items. Multi-stakeholder optimization [35], [36] argues for context-aware definitions (e.g., exposure vs. acceptance parity), while text-based relevance features (e.g., TF–IDF) can help ranking [37] but do not, on their own, guarantee equity for historically marginalized authors.

We focus on a *post-review* neural re-ranking stage with a differentiable fairness loss that handles *intersectional* attributes. The synthetic analysis provides a controlled test of the hypothesis that increasing fairness regularization improves inclusion with limited utility cost—and reveals regime-dependent sweet spots. The original-data validation then asks whether these settings hold under real constraints shaped by industry–academia participation, collaboration networks, and conference practices [19], [20], [22], [24], [25]. This two-stage path converts a simple, testable idea into a practical recipe for equity-preserving paper selection.

III. Methodology

Our approach tackles demographic biases in conference data to enforce fairness post-review: we begin with a simple, controlled hypothesis tested on synthetic datasets, then validate its robustness on original conference data. The design therefore emphasizes both (i) tractability in controlled regimes and (ii) fidelity to the complex realities of peer review. The backbone of our approach is a lightweight Multilayer Perceptron (MLP) trained with an integrated fairness loss, as illustrated in Fig. 1. This choice reflects two guiding principles: (1) expose and alleviate demographic disparities instead of concealing them, and (2) use a simple, transparent neural architecture that harmonizes fairness and utility while remaining interpretable enough to trace effects across bias regimes.

A. Data Collection and Pre-processing

a) Synthetic stage (hypothesis testing): Synthetic datasets are constructed with tunable demographic distributions (fair, moderate, high bias). These provide a sandbox where the effect of fairness regularization λ can be precisely measured. By varying the demographic skew, we can test the hypothesis that fairness constraints are most effective when initial bias is high, and that "sweet spot" λ values differ by attribute. Table I summarizes the demographic distributions for each protected group across the fair, moderate, and high synthetic bias scenarios.

b) Original stage (validation): To evaluate transfer to real contexts, we utilize datasets from SIGCHI 2017, DIS 2017, and IUI 2017 [32]. These venues naturally exhibit systemic imbalances in demographics and prestige, reflecting the challenges of real peer review. Rather than discarding these biases, we preserve them to observe how fairness regularization reshapes outcomes in situ. Table II reports the participation rates of protected groups (gender, race, country) in the three conferences (SIGCHI, DIS, IUI).

c) Processing pipeline: At both stages, the dataset is curated at paper and author levels. Every paper record includes a title, author list, and conference label (1 = IUI, 2 = DIS, 3 = SIGCHI). Author records capture demographic variables (gender, race, nationality, career stage). Real-world datasets—particularly those drawn from academic conference submissions—often contain

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I. Demographic Distribution in Synthetic Datasets Across Bias Scenarios For Protected Groups

Scenario	Gender (%)	Race (%)	Country (%)
Fair	48.80	51.50	52.30
Moderate	28.10	28.90	31.50
High	8.50	9.70	10.00
Average	28.47	30.03	31.27

latent biases that mirror systemic imbalances in the scholarly community (e.g., underrepresentation of certain demographics). We utilize datasets from conferences that naturally reflect systemic disparities (e.g., skewed demographics). Instead of eliminating such biases, our objective is to *recognize and rectify* them.

TABLE II. DEMOGRAPHIC PARTICIPATION FROM PROTECTED GROUPS IN THREE CONFERENCES.

Conference	Gender (%)	Race (%)	Country (%)
SIGCHI	41.88	6.84	21.94
DIS	65.79	35.09	24.56
IUI	43.75	51.56	39.06
Average	50.47	31.16	28.52

1) Data Description: The resulting dataset combines both paper-level and author-level perspectives. *Overall* refers to all submissions, while *Selected* refers to those recommended by Fair-PaperRec. SIGCHI 2017 papers act as a benchmark for high-impact work, offering a ground truth for quality. Author records include career stages (student, postdoc, faculty, industry), enabling h-index weighting that reflects differences in scholarly maturity.

2) Data Pre-processing: Several preprocessing steps were undertaken to prepare the dataset for training:

Categorical Encoding: Gender, Country, and Race are subjected to one-hot encoding. Gender is binary (0 = male, 1 = female), Country is categorized as *developed* or *underdeveloped*, and Race comprises {White, Asian, Hispanic, Black}, with Hispanic and Black designated as protected groups (Table II). *Normalization:* Numerical attributes (e.g., h-index) employ min-max scaling for consistent magnitude. *Training and Validation Division:* An 80%/20% stratified division guarantees equitable distribution of labels and protected attributes in both subsets.

B. Problem Definition

This study develops a *fairness-aware paper recommendation system* that ensures demographic parity with respect to authors' race and country while preserving high academic standards. We frame acceptance decisions as a *recommendation* task, where *conference organizers (users)* seek to select from *530 papers (items)* spanning SIGCHI, DIS, and IUI. Each paper (item) includes an *h-index* for quality, demographic data (race, country), and a conference rating.

Our approach enforces fairness constraints on race and country independently, excluding *gender* due to its relatively balanced distribution in the dataset (Table II). Preliminary analysis further showed minimal disparity between accepted and rejected papers with respect to gender; because our fairness framework penalizes deviations from the observed baseline, incorporating gender would have produced negligible changes in the fairness–utility trade-offs examined here. To maintain clarity of exposition and focus on attributes with substantially higher disparity, we therefore restricted fairness constraints to race and country. Nevertheless, gender remains an important dimension of equity, and future work will extend Fair-PaperRec to gender-aware and multi-attribute fairness in larger or more imbalanced datasets.

Let D represent the dataset of submitted papers, where each paper $p \in D$ is associated with a set of features X_p (e.g., race, country, h-index) and a target variable y_p indicating acceptance (1) or rejection (0). The *race* attribute R_p and *country* attribute C_p are the protected attributes.

We aim to optimize a predictive model $f : X_p \rightarrow \hat{y}_p$ that minimizes the following objective function:

$$\min_f \left(\mathcal{L}(f(X_p), y_p) + \lambda \cdot \mathcal{L}_{\text{fairness}}(f, D)\right) \tag{1}$$

Here, $\mathcal{L}(f(X_p), y_p)$ is the *prediction loss* (e.g., Binary Cross-Entropy Loss), $\mathcal{L}_{\text{fairness}}(f, D)$ is the *fairness loss*, penalizing deviations from demographic parity across race and country, and λ is a hyperparameter that balances the trade-off between prediction accuracy and fairness.

Algorithm 1. FAIR-PAPERREC LOSS FUNCTION.

Input: Model M, Epochs E, Batch size B, Data D, Protected attributes A, Hyperparameter λ
Output: Trained Model M
Initialize Model M
for each $e \in E$ **do**
  Shuffle Data D
  for each batch $\{(X, Y)\} \in D$ with size B **do**
    Predict $\hat{Y} \leftarrow M(X)$
    Calculate Loss:
      $L_{\text{prediction}} \leftarrow \text{PredictionLoss}(Y, \hat{Y})$
      $L_{\text{fairness}} \leftarrow \text{FairnessLoss}(A, \hat{Y})$
    Calculate Total Loss:
      $L_{\text{total}} \leftarrow \lambda \cdot L_{\text{fairness}} + L_{\text{prediction}}$
    Compute gradients $\nabla L_{\text{total}} \leftarrow \frac{\partial L_{\text{total}}}{\partial M}$
    Update Model parameters: $M \leftarrow M - \alpha \nabla L_{\text{total}}$
  end for
end for

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

C. Demographic Parity

We aim to ensure that the probability of a paper being accepted is independent of the protected attributes:

$$P(\hat{y}_p = 1 \mid R_p = r) = P(\hat{y}_p = 1), \quad \forall r \in \text{Race}$$

$$P(\hat{y}_p = 1 \mid C_p = c) = P(\hat{y}_p = 1), \quad \forall c \in \text{Country}$$

Utilizing these equations ensures that the papers authored by individuals from different races and countries have an equal probability of acceptance.

D. Fairness Loss

The fairness loss from the objective function in Equation (1) is constructed to minimize statistical parity differences between the protected and non-protected groups:

$$\mathcal{L}_{\text{fairness}} = \left(P(\hat{y}_p = 1 \mid G_{\text{p}}) - P(\hat{y}_p = 1 \mid G_{\text{np}})\right)^2 \quad (2)$$

Here, $P(\hat{y}_p = 1 \mid G_{\text{p}})$ denotes the acceptance probability for the protected group and $P(\hat{y}_p = 1 \mid G_{\text{np}})$ is the acceptance probability for the non-protected group.

E. Combined Fairness Loss

Furthermore, we define a combined fairness loss to minimize statistical parity differences across race and country attributes between the protected and unprotected groups, as shown in Equation (3).

$$\begin{aligned}\mathcal{L}_{\text{fairness}} = W_r &\left(\frac{1}{N_r}\sum_{p\in G_r}\hat{y}_p - \frac{1}{N}\sum_{p=1}^{N}\hat{y}_p\right)^2 \\ &+ W_c\left(\frac{1}{N_c}\sum_{p\in G_c}\hat{y}_p - \frac{1}{N}\sum_{p=1}^{N}\hat{y}_p\right)^2\end{aligned} \quad (3)$$

G_{r} and G_{c} denote the race and country groups, respectively. N_r and N_c are the number of papers in each group and weights W_r and W_c reflect group distributions.

F. Total Loss

The total loss is the combination of prediction and fairness losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{prediction}} + \lambda \cdot \mathcal{L}_{\text{fairness}}$$

G. Constraints and Considerations

We assess fairness by training our model separately on *race* and *country*, as well as jointly on both attributes to evaluate selection fairness across multiple dimensions.

a) Exclusion of Protected Attributes: Race R_p and country C_p are excluded from the input feature set X_p to mitigate direct bias amplification. To achieve joint fairness, both attributes are omitted during training, preventing the model from learning acceptance outcomes influenced by race or country.

b) Indirect Bias Mitigation: A fairness loss promotes demographic parity, addressing indirect biases associated with features related to race or country. The model maintains neutrality by penalizing selection disparities, even in the absence of protected attributes.

c) Scalability: The MLP architecture is deliberately simple, enabling reproducibility across venues and bias levels. Synthetic results provide controlled insight into when fairness is most beneficial; original results demonstrate external validity. Together, they build a methodological bridge from a simple hypothesis ("fairness regularization should help most under bias") to a powerful conclusion ("in real conferences, tuning λ uncovers under-selected, high-quality work without compromising rigor").

IV. Model Overview

To achieve demographic parity while preserving quality in paper selection, we present a MLP-based neural network (see Fig. 1), explicitly engineered to balance the trade-off between fairness and accuracy. It illustrates the correlations between input features, like author demographic attributes and paper quality, while alleviating biases during selection.

A unique fairness loss function was employed to ensure equity, imposing penalties on the model for substantial differences in selection rates between protected and non-protected groups. This loss function is integrated with the conventional prediction loss to attain a balance between diversity and accuracy; the algorithm is shown in Algorithm 1.

The acceptance probabilities for submitted papers are generated by the MLP, which are subsequently ranked to guarantee that the final selection meets both quality and fairness objectives. By selecting top papers according to these probabilities, we ensure equal representation of authors from both protected and non-protected groups while upholding the requisite standard of academic excellence.

Algorithm 2. FAIRNESS-AWARE PAPER SELECTION MECHANISM.

Input: Dataset D, Model M, Number of Accepted Papers N_a, Total Papers N_t
Output: Selected Papers P_{selected}
Initialize: $P_{\text{selected}} \leftarrow \emptyset$
Step 1: Apply trained model M to the entire dataset D
for each paper $p \in D$ **do**
    Compute acceptance probability: $\hat{y}_p \leftarrow M(p)$
end for
Step 2: Rank all papers p by acceptance probability $\hat{y}_p$
Sort D in descending order of $\hat{y}_p$
Step 3: Select top N_a papers:
    $P_{\text{selected}} \leftarrow \{p \mid \hat{y}_p \geq \hat{y}_{(N_a)}\}$
Step 4: Ensure Fairness Constraints
Return P_{selected}

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE III. RELATIVE GAIN (%) FOR MACRO, MICRO, AND UTILITY METRICS ACROSS $\lambda \in [1, 10]$ FOR COUNTRY AND RACE FEATURES UNDER DIFFERENT BIAS LEVELS.

		Country Feature			Race Feature		
Bias Level	**λ**	**Macro Gain (%)**	**Micro Gain (%)**	**Utility Gain (%)**	**Macro Gain (%)**	**Micro Gain (%)**	**Utility Gain (%)**
High	1	-18.03	-4.17	-2.98	**15.03**	20.18	8.51
	2.5	-3.93	-1.65	-1.28	12.44	18.46	2.55
	3	-1.18	3.22	5.96	12.11	**21.10**	**15.32**
	5	-0.30	-2.61	8.09	10.12	16.84	14.47
	10	**4.98**	**9.22**	**10.21**	5.06	14.20	5.53
Moderate	1	2.90	**10.44**	**15.32**	3.00	**10.62**	11.06
	2.5	-1.11	-2.83	4.68	1.82	4.66	8.09
	3	**5.04**	3.97	7.23	**3.96**	9.54	**13.62**
	5	-3.15	-6.15	0.43	4.39	10.55	11.06
	10	2.29	0.49	-0.85	0.83	3.05	7.66
Fair	1	**5.35**	**7.26**	**20.43**	-6.51	0.79	-4.26
	2.5	0.51	0.02	8.94	-0.28	-0.30	11.06
	3	3.81	2.01	-0.85	-8.96	-12.78	8.09
	5	2.31	1.43	3.83	-2.12	-4.76	**12.77**
	10	3.93	2.12	10.64	**1.37**	-1.78	4.68

A. Selection Mechanism

The model calculates acceptance probabilities for all submitted papers after training. After calculating acceptance odds, the algorithm ranks candidate papers. This rating phase ensures that underrepresented groups are represented in the final admission decisions. Representing this as a suggestion list preserves the peer-review process and corrects residual biases. Algorithm 2 selects the best papers based on probability, ensuring fairness and preserving the desired number of accepted papers.

- *Prediction Aggregation:* The trained MLP model is applied to the entire dataset to obtain predicted acceptance probabilities $\hat{y}_p$ for each paper.
- *Ranking:* Papers are ranked in descending order based on their predicted probabilities.
- *Selection:* The papers with the highest predicted probabilities are selected for acceptance, ensuring that the total number of selected papers matches the required acceptance quota.

Mathematically, the selection process is represented as:

$$\text{Selected Papers} = \left\{p \in D \mid \hat{y}_p \geq \hat{y}_{(N_a)}\right\}$$

Here, $\hat{y}_{(N_a)}$ is the N_a-th highest predicted probability in the set $\{\hat{y}_p \mid p \in D\}$ while N_a is the total number of accepted papers and N_t is the total number of submitted papers, where $N_a \leq N_t$.

This approach ensures that the selection process is both informed by the model's predictions and constrained to uphold demographic parity, fostering an equitable and meritocratic paper selection environment.

V. Evaluation and Experiments

This section presents the experimental evaluation of our proposed Fair-PaperRec model on the chosen datasets. First, we test our hypothesis in controlled synthetic settings, then we validate it with original conference data. This design allows us to ask not only whether fairness interventions work, but also *when* and *why* they are most effective.

To guide this exploration, we pose three research questions:

- *RQ1:* How do fairness constraints affect the overall quality (utility) of recommended papers, as measured by author h-index?
- *RQ2:* Does handling race and country as separate protected attributes differ from treating them jointly in terms of fairness outcomes and selection decisions?
- *RQ3:* How do varying weight assignments to multiple protected attributes (race and country) influence the trade-off between fairness and utility?

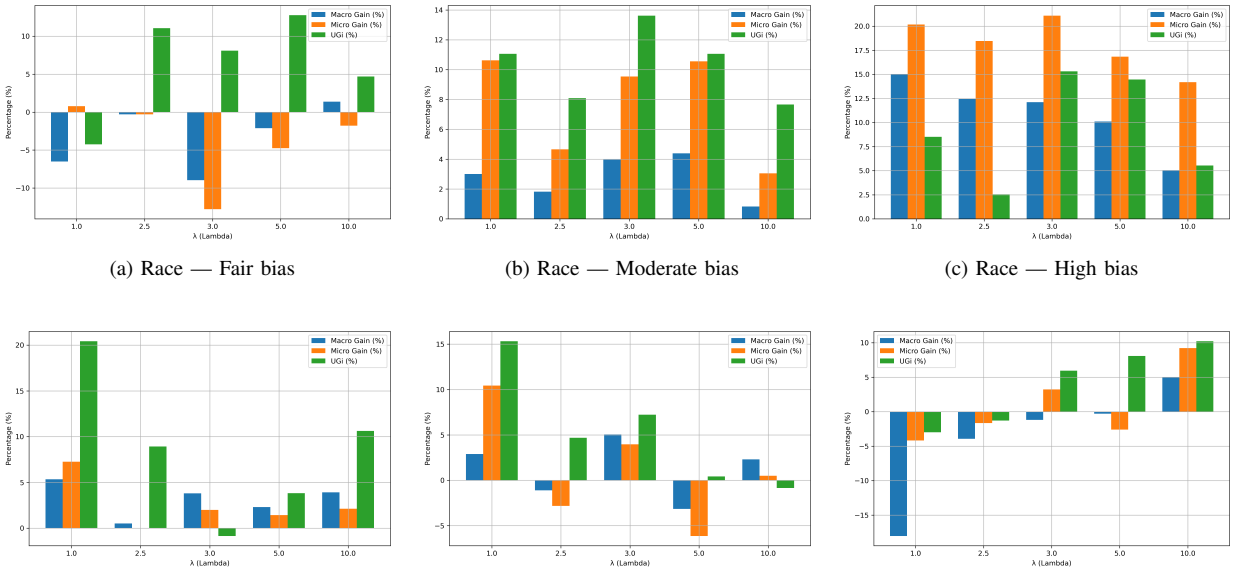
These questions move us along the synthetic-to-original trajectory, from controlled hypothesis testing to real-world validation. RQ1 begins in the synthetic environment, where we test the basic hypothesis that fairness regularization (λ) improves inclusion with limited cost to utility. RQ2 and RQ3 extend to the real-world stage, asking whether the same patterns hold under authentic demographic distributions and multi-attribute interactions.

A. Experimental Setting

We evaluate Fair-PaperRec across both synthetic datasets (fair, moderate, and high-bias regimes) and original conference datasets. This dual evaluation ensures that our findings generalize from controlled testbeds to complex real-world settings where structural factors (prestige, career stage, institutional affiliation) interact with demographics.

Each experiment is conducted five times with independent random seeds. We report mean values with *standard deviations* to highlight stability across runs, ensuring

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) Race — Fair bias (b) Race — Moderate bias (c) Race — High bias

(d) Country — Fair bias (e) Country — Moderate bias (f) Country — High bias

Figure 2. Relative gains (%) for Macro, Micro, and Utility across fairness regularization strengths ($\lambda \in [1, 10]$). Top row: Race; bottom row: Country.

that observed peaks or trade-offs are not artifacts of stochastic training.

1) Implementation Details: All experiments are implemented in `PyTorch` and executed on a high-performance server equipped with two NVIDIA Quadro RTX 4000 GPUs. The Fair-PaperRec model is a two-hidden-layer MLP using Rectified Linear Unit (ReLU) activations and Batch Normalization, culminating in a sigmoid output layer that produces acceptance probabilities.

TABLE IV. DISTRIBUTION OF RECOMMENDED PAPERS FROM EACH TIER FOR SYNTHETIC DATASET.

Label	Country	Race	Multi-Fair
Top-tier	87.12%	89.30%	91.41%
Mid-tier	6.45%	6.87%	5.64%
Low-tier	6.43%	3.83%	2.95%
# Papers	280	280	280

We train for 50 epochs using the Adam optimizer (learning rate = 0.001). Early stopping is applied if validation performance does not improve over 10 consecutive epochs. The fairness regularization parameter λ is swept across a range of values (1–10) to trace fairness–utility trade-offs. Datasets are split 80/20 (training/validation) using stratified sampling, preserving both label and demographic distributions.

This setup allows us to simulate the "synthetic hypothesis-testing stage" (controlled bias ratios) and then apply the same procedure in the "original validation stage" (real demographic skew).

TABLE V. DISTRIBUTION OF RECOMMENDED PAPERS FROM EACH CONFERENCE.

Label	Country	Race	Multi-Fair
SIGCHI	92.02%	92.00%	92.02%
DIS	4.84%	7.69%	7.40%
IUI	3.14%	0.31%	0.56%
# Papers	351	351	351

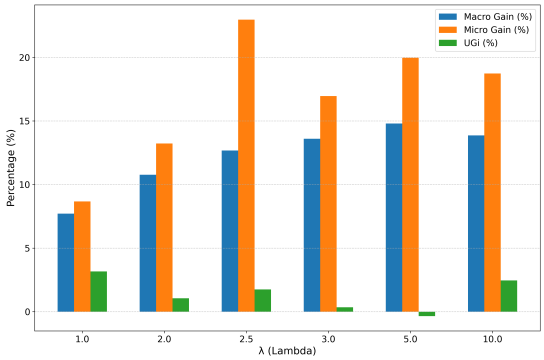


Figure 3. Comparison of Macro and Micro Gains for Country Across Different Fairness Configurations.

2) Baseline: We compare our model against a baseline demographic-blind model, which is a conventional (MLP) model that prioritizes quality and ignores fairness constraints. This model selects the original list of papers chosen by the SIGCHI 2017 program committee. By contrasting it with Fair-PaperRec, we identify λ values where fairness gains are achieved without utility loss (addressing RQ1). In original settings, the baseline rep-

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

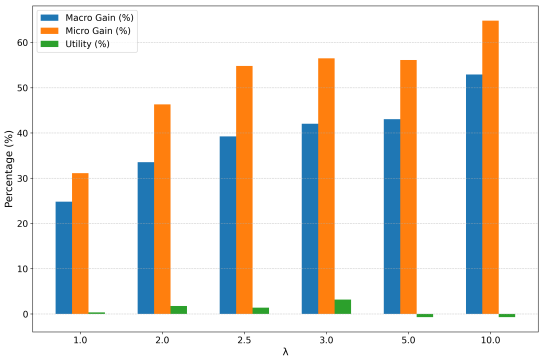


Figure 4. Comparison of Macro and Micro Gains for Race Across Different Fairness Configurations.

resents the actual peer-review outcomes; improvements over it demonstrate the practical value of fairness interventions in real conferences (addressing RQ2 and RQ3).

3) Parameters: A hyperparameter λ is used to control the trade-off between prediction accuracy and fairness. Higher values emphasize fairness more strongly.

The weights W_{c} and W_{r}, respectively, denote the weighting factors assigned to the country and race attributes in the fairness loss function, as shown in Equation (3).

B. Evaluation Metrics

Diversity is assessed at both the *paper level* and the *author level*. In particular:

- *Macro Gain* represents the percentage increase in the diversity of each feature within the selected papers compared with the baseline, assessing the overall representation of protected groups.
- *Micro Gain* is the percentage increase in the diversity of each feature among authors of the selected papers, providing a more detailed perspective on inclusivity.
- A *Diversity Gain* [32] further normalizes these macro-level changes in Equation (4), capping each feature at 100 to avoid any single attribute skewing the total.
- The *F - measure* [32] Equation (5) then combines this diversity improvement with the resulting utility, offering a harmonic balance between fairness gains and paper quality.

Table VII summarizes the fairness, diversity, and utility metrics used in our evaluation, including Macro Gain, Micro Gain, Diversity Gain, and the F-measure.

To ensure that diversity enhancements do not compromise scholarly quality, we assess *Utility Gain* (UG_i) using a weighted h-index that reflects an author's career stage (professor, associate professor, lecturer, postdoctoral researcher, or graduate student). While the h-index offers a convenient proxy for scholarly influence, it also encodes seniority and field-specific publication practices that may introduce their own biases. For this reason, we interpret utility gain cautiously, not as an absolute measure of "quality," but as a stability indicator demonstrating whether fairness interventions disproportionately select low impact papers. By comparing the resulting utility values against the baseline distribution, we evaluate whether fairness-aware re-ranking maintains, improves, or undermines the overall quality profile of accepted papers.

$$D_G = \frac{\sum_{i=1}^{n} \min(100, \text{Macro Gain}_{G_i})}{n} \tag{4}$$

$$F = 2 \times \frac{D_G \times (100 - UG_i)}{D_G + (100 - UG_i)} \tag{5}$$

C. Interpretation of the Results

We evaluate the fairness regularization parameter (λ) across values 1–10 to examine its impact on fairness (macro and micro gains) and quality (Utility Gain, UGi). The results span both synthetic datasets and original conference datasets. This two-stage design allows us to first test the hypothesis under controlled conditions and then validate it using real-world peer review data.

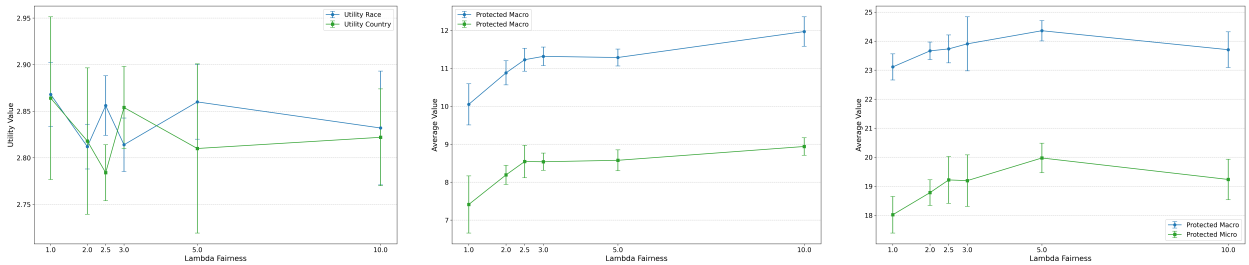
a) RQ1: Effect of fairness on utility: Fig. 2 (synthetic) and Figs. 4–3 (original) show that increasing λ generally improves macro and micro diversity, while the trajectory of utility differs across settings. Detailed gain calculations are reported in Table III.

- In synthetic *high-bias* settings, we observe a "sweet spot" ($\lambda \approx 3$) where both fairness and utility peak together. This supports our hypothesis that correcting bias can uncover under-selected, high-quality work.
- In synthetic *moderate/fair* settings, small λ values act as a useful regularizer, but large values λ overcorrect, harming utility.
- In original data, utility is more stable for race (flat or mildly positive trend) but fluctuates for country, with error bars indicating greater uncertainty at higher λ. This confirms the trade-off: fairness gains are achievable, but utility must be monitored carefully.

TABLE VI. GAIN CALCULATIONS FOR COUNTRY AND RACE FEATURES WITH UTILITY GAIN (UG_{i}).

	Country Feature			Race Feature		
λ	Macro Gain (%)	Micro Gain (%)	UG_{i} (%)	Macro Gain (%)	Micro Gain (%)	UG_{i} (%)
1	7.71	8.67	3.16	24.81	31.11	0.35
2	10.77	13.23	1.05	33.54	46.30	1.75
2.5	**12.67**	**22.96**	**1.75**	39.25	54.81	1.40
3	13.60	16.96	0.35	**42.03**	**56.48**	**3.16**
5	14.80	19.97	-0.35	43.04	56.11	-0.70
10	13.86	18.73	2.46	52.91	64.81	-0.70

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org



(a) Utility Gain. (b) Macro/Micro for Race. (c) Macro/Micro for Country.

Figure 5. Comparison of gains across different fairness configurations.

b) RQ2: Attribute-specific effects (race vs. country): The optimal λ differs between attributes. Race requires larger λ values (around 3) to balance fairness and utility, reflecting the higher initial disparity ratios of racial minorities. Country achieves balance at lower λ values (around 2.5). Synthetic experiments revealed the same pattern: the more skewed the initial distribution, the stronger the regularization required. This validates our hypothesis that fairness policies must be tailored to group-specific disparities. A summary of these trade-offs for different λ values and weight assignments is provided in Table VII.

c) RQ3: Multi-attribute and weighting trade-offs: Experiments that combine race and country with varying weights show that emphasizing one attribute can reduce gains for the other. On synthetic data, joint optimization occasionally diluted improvements compared to single-attribute runs. On original data, this was visible in Micro diversity for country, which exhibited more volatility than race under joint fairness. As summarized in Fig. 5, these configurations reveal how different fairness weightings reshape the balance between utility and diversity, making the trade-offs between race-focused, country-focused, and multi-attribute fairness more explicit.

d) Macro vs. Micro Dynamics: Across both synthetic and original settings, macro diversity measures were consistently more responsive to fairness regularization than micro measures. Macro gains rose steadily with λ, especially for race, whereas micro gains were more variable (particularly for country). This suggests that fairness interventions are more effective at improving overall group-level representation than at equalizing outcomes for individual authors.

e) Conference-level distribution: The detailed gain calculations for each λ are shown in Table VI, which reports the breakdown of accepted papers across SIGCHI, DIS, and IUI under different fairness constraints. SIGCHI dominates acceptance at approximately 92% regardless of λ, reflecting its strong baseline prestige. However, fairness constraints modestly increase the share of DIS and IUI papers, broadening representation without destabilizing the overall distribution. This indicates that fairness-aware re-ranking can diversify the conference portfolio while preserving dominant trends. For context, the original conference-level distribution of accepted papers is summarized in Table V. In the synthetic stage, the baseline distribution is configured to mimic this original selection profile under controlled bias levels; the corresponding breakdown is reported in Table IV.

D. Robustness & Sensitivity

For completeness, we report seed variability (mean$\pm$std over multiple random seeds) and, when space allows, confidence intervals around the λsweeps to demonstrate that peaks (e.g., Race–High at $\lambda \approx 3$) are stable rather than artifacts of stochasticity.

E. Ablation Study: Multi-Demographic Fairness

The objective of our ablation study was to evaluate the model's performance when optimizing fairness across multiple demographic attributes simultaneously, specifically with respect to both *country* and *race*. This ablation was conducted to address *RQ3*, which explores the impact of varying fairness weights for each attribute when multiple fairness attributes are considered together.

To ensure fairness, we removed these attributes from the input space, preventing the model from learning direct associations between them and the paper acceptance decisions. Instead, demographic parity loss was computed for each attribute during training, capturing deviations from fairness. The parity losses for both country and race were combined by assigning weights: W_c for country and W_r for race, with the initial weights set to $W_c = 0.68$ and $W_r = 0.32$, reflecting the distribution of protected groups.

To further explore the model's behavior and answer *RQ3*, we varied these weights, first increasing W_c while keeping W_r constant and then increasing W_r while keeping W_c fixed. Additionally, we experimented with different values of the fairness regularization parameter λ, which controls the trade-off between fairness and utility. These experiments allowed us to observe how different weight configurations and fairness constraints

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE VII. GAIN CALCULATIONS FOR COUNTRY AND RACE FEATURES WITH UTILITY GAIN.

λ	Weights	Country Feature		Race Feature		UG_i (%)	Avg. D_G (%)	Avg. F (%)
		Macro Gain (%)	Micro Gain (%)	Macro Gain (%)	Micro Gain (%)			
1	($W_r = 0.32,\ W_c = 0.68$)	6.17	6.34	30.51	46.30	3.16	44.66	53.71
	($W_r = 1,\ W_c = 2$)	6.73	9.15	-0.25	0.37	2.81	6.48	13.77
	($W_r = 2,\ W_c = 1$)	7.43	11.43	12.91	16.11	3.16	25.63	40.36
2	($W_r = 0.32,\ W_c = 0.68$)	**13.60**	**24.43**	30.51	42.22	**4.21**	55.38	68.47
	($W_r = 1,\ W_c = 2$)	5.24	6.88	15.45	17.96	0.70	20.69	21.58
	($W_r = 2,\ W_c = 1$)	8.36	12.86	39.49	54.26	1.75	26.31	21.58
2.5	($W_r = 0.32,\ W_c = 0.68$)	8.63	17.33	36.58	50.37	2.46	56.46	66.31
	($W_r = 1,\ W_c = 2$)	9.89	14.00	30.63	46.30	2.81	40.52	62.09
	($W_r = 2,\ W_c = 1$)	9.60	17.11	42.53	56.48	1.40	59.25	69.98
3	($W_r = 0.32,\ W_c = 0.68$)	7.15	11.42	39.49	53.89	1.40	55.98	63.45
	($W_r = 1,\ W_c = 2$)	10.16	21.17	33.29	43.89	0.70	43.45	47.63
	($W_r = 2,\ W_c = 1$)	9.60	18.35	42.53	55.37	2.81	61.90	47.63
5	($W_r = 0.32,\ W_c = 0.68$)	10.80	19.38	**45.82**	**58.52**	0.70	**65.09**	**72.92**
	($W_r = 1,\ W_c = 2$)	4.69	3.88	33.92	40.19	0.35	38.61	15.73
	($W_r = 2,\ W_c = 1$)	7.43	11.90	39.49	52.96	5.26	52.26	15.73
10	($W_r = 0.32,\ W_c = 0.68$)	9.60	18.34	42.53	55.37	1.40	62.92	70.89
	($W_r = 1,\ W_c = 2$)	7.43	13.91	24.94	25.19	4.91	32.37	34.88
	($W_r = 2,\ W_c = 1$)	7.43	11.72	35.44	47.41	-4.21	40.53	34.88

influenced the model's ability to achieve demographic fairness while maintaining utility and the quality of selected papers.

The results of the ablation study, shown in Table VII, reveal that at $\lambda = 1$, assigning equal weights to both race and country ($W_r = 0.32, W_c = 0.68$) produced significant gains for race, with a Macro Gain of 30.51% and a Micro Gain of 46.3%, while country showed relatively smaller improvements (6.17% and 6.34%, respectively). However, when the weight for country was increased ($W_c = 2 \times 0.68$), diversity gains for race dropped sharply, with a negative macro gain (-0.25%), while country experienced slight improvements. Conversely, increasing the weight for race ($W_r = 2{\times}0.32$) resulted in improved diversity for both race and country, indicating that assigning more weight to race enhances diversity for both attributes to some degree.

At $\lambda = 2.5$, the model achieved the best balance between diversity and utility. Equal weights for race and country yielded macro and micro gains of 36.58% and 50.37% for race, and 8.63% and 17.33% for country, with a low utility loss of 2.46%. This suggests that $\lambda = 2.5$ is optimal for balancing fairness and utility. As λ increases further, race diversity continues to improve (reaching 45.82% Macro Gain at $\lambda = 5$), but at the cost of decreasing utility. The different optimal λ values for race and country suggest that disparity ratios impact how fairness constraints should be weighted, with race requiring a higher λ due to its higher disparity ratio. This leads to greater race diversity gains at higher λ values, whereas country achieves optimal results at moderate λ values, such as 2.5.

These findings directly address *RQ3*, demonstrating that fairness weights must be carefully calibrated for each protected attribute. Assigning greater weight to race tends to improve diversity for both race and country, whereas increasing the weight for country may result in reduced fairness for race. The optimal balance between fairness and utility is achieved when fairness weights and λ values are adjusted based on the unique disparity ratios of each attribute.

VI. Conclusion and Future Work

We began our experiment with a hypothesis tested on synthetic data: a lightweight fairness regularizer, controlled by λ, could improve demographic inclusion without compromising quality. Applying the same regularization to the original conference data confirmed and strengthened this insight: fairness regularization consistently improved representation, particularly for race, while leaving utility largely stable or mildly positive.

These results highlight three key lessons. First, fairness effects depend on the degree of underlying disparity: high-bias settings benefit from stronger regularization, while near-fair regimes require small adjustments to avoid over-correction. Second, attributes vary in sensitivity: race often demands more intervention than country, and multi-attribute optimization requires careful balancing. Third, in strongly biased systems, fairness and quality are not in conflict; debiasing uncovers underselected, high-quality work.

At the same time, our approach has limits. We rely on a simple MLP with post-hoc fairness loss, excluding protected attributes from the input but not explicitly modeling causal pathways or reviewer dynamics. Extending this framework with causal inference, graph-based architectures capturing author–institution–topic relations, or generative approaches such as VAEs could deepen bias mitigation while preserving interpretability. Broader definitions of fairness, adaptive tuning of λ and attribute weights, and human-in-the-loop evaluation are promising directions. Embedding fairness interventions into real-world workflows will require robustness to distribution shifts, transparency in reporting, and governance

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

mechanisms that account for privacy and accountability. Together, these steps can transform Fair-PaperRec from a research prototype into a practical, equitable tool for the future of scholarly peer review.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) under Award number OIA-1946391, Data Analytics that are Robust and Trusted (DART).

REFERENCES

[1] U. A. Oyshi and S. Gauch, "Fair learning for bias mitigation and quality optimization in paper recommendation," in *Proceedings of the 17th International Conference on Information, Process, and Knowledge Management (eKNOW 2025)*, ThinkMind Digital Library, Nice, France: IARIA Press, May 2025, pp. 43–50, ISBN: 978-1-68558-272-2.

[2] A. Tomkins, M. Zhang, and W. D. Heavlin, "Reviewer bias in single- versus double-blind peer review," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 48, pp. 12 708–12 713, 2017. DOI: 10.1073/pnas.1707323114.

[3] E. Shmidt and B. Jacobson, "Double-blind reviews: A step toward eliminating unconscious bias," *Clinical and Translational Gastroenterology*, vol. 13, no. 1, e00443, 2022. DOI: 10.14309/ctg.0000000000000443.

[4] V. P. Giannakakos, T. S. Karanfilian, A. D. Dimopoulos, and A. Barmettler, "Impact of author characteristics on outcomes of single- versus double-blind peer review: A systematic review of comparative studies in scientific abstracts and publications," *Scientometrics*, vol. 130, pp. 399–421, 2025. DOI: 10.1007/s11192-024-05213-x.

[5] C. Mebane, "Double-blind peer review is detrimental to scientific integrity," *Environmental Toxicology and Chemistry*, vol. 44, pp. 318–323, 2025. DOI: 10.1093/etojnl/vgae046.

[6] L. Bauersfeld, A. Romero, M. Muglikar, and D. Scaramuzza, "Cracking double-blind review: Authorship attribution with deep learning," *PLoS ONE*, vol. 18, no. 6, e0287611, Jun. 2023. DOI: 10.1371/journal.pone.0287611.

[7] N. B. Shah, "The role of author identities in peer review," *PLOS ONE*, vol. 18, no. 6, e0286206, Jun. 2023. DOI: 10.1371/journal.pone.0286206.

[8] J. Huber et al., "Nobel and novice: Author prominence affects peer review," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 41, e2205779119, 2022. DOI: 10.1073/pnas.2205779119.

[9] E. Frachtenberg and K. McConville, "Metrics and methods in the evaluation of prestige bias in peer review: A case study in computer systems conferences," *PLoS ONE*, vol. 17, e0264131, 2022. DOI: 10.1371/journal.pone.0264131.

[10] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin, "Bias in peer review," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 1, pp. 2–17, 2013. DOI: 10.1002/asi.22784.

[11] C. L. Goues et al., "Effectiveness of anonymization in double-blind review," *Communications of the ACM*, vol. 61, pp. 30–33, 2017. DOI: 10.1145/3208157.

[12] J. Bobadilla, R. Lara-Cabrera, A. González-Prieto, and F. Ortega, "DeepFair: Deep learning for improving fairness in recommender systems," *Information Processing & Management*, vol. 58, no. 3, p. 102 547, May 2021. DOI: 10.1016/j.ipm.2021.102547.

[13] Y. Peng, X. Qian, and W. Song, "A re-ranking approach for two-sided fairness on recommendation systems," in *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, May 2023, pp. 312–316. DOI: 10.1145/3603781.3603836.

[14] M. Morik, A. Singh, J. Hong, and T. Joachims, "Controlling fairness and bias in dynamic learning-to-rank," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, 2020, pp. 267–276. DOI: 10.1145/3340531.3412875.

[15] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, *Data decisions and theoretical implications when adversarially learning fair representations*, FAT/ML 2017 Workshop, Accessed: 2025-12-01, 2017.

[16] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[17] S. Abimbola, M. Pai, P. Nguyen, T. Greenhalgh, and E. H. Shadid, "How to address the geographical bias in academic publishing," *BMJ Global Health*, vol. 8, no. Suppl 2, e013111, Dec. 2023. DOI: 10.1136/bmjgh-2023-013111.

[18] W. M. Williams and S. J. Ceci, "National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 17, pp. 5360–5365, Apr. 2015. DOI: 10.1073/pnas.1418878112.

[19] T. Hagendorff and K. Meding, "Ethical considerations and statistical analysis of industry involvement in machine learning research," *AI & SOCIETY*, vol. 38, pp. 35–45, 2023. DOI: 10.1007/s00146-021-01284-z.

[20] M. Bosten and B. Kleinberg, *Conflicts of interest in published nlp research 2000–2024*, ArXiv, vol. abs/2502.16218, Accessed: 2025-12-01, 2025. DOI: 10.48550/arXiv.2502.16218.

[21] I. Kotseruba, M. Papagelis, and J. K. Tsotsos, *Industry and academic research in computer vision*, ArXiv, vol. abs/2107.04902, Accessed: 2025-12-01, 2021. DOI: 10.48550/arXiv.2107.04902.

[22] L. Liang, H. Zhuang, J. Zou, and D. E. Acuna, *The complementary contributions of academia and industry to AI research*, ArXiv, vol. abs/2401.10268, Accessed: 2025-12-01, 2024. DOI: 10.48550/arXiv.2401.10268.

[23] M. Färber and L. Tampakis, "Analyzing the impact of companies on AI research based on publications," *Scientometrics*, vol. 129, pp. 31–63, 2024. DOI: 10.1007/s11192-023-04867-3.

[24] J. L. Ortega, "Influence of co-authorship networks in the research impact: Ego network analyses from Microsoft Academic Search," *Journal of Informetrics*, vol. 8, no. 3, pp. 728–737, 2014. DOI: 10.1016/j.joi.2014.07.001.

[25] M. Y. Vardi, "Conferences vs. journals in computing research," *Communications of the ACM*, vol. 52, no. 5, p. 5, 2009. DOI: 10.1145/1506409.1506410.

[26] B. Meyer, *Reversing the fossilization of computer science conferences*, Communications of the ACM, BLOG@CACM, Accessed: 2025-12-01, Apr. 2025.

[27] J. M. Fernandes, A. Costa, and P. Cortez, "Author placement in computer science: A study based on the careers of ACM fellows," *Scientometrics*, vol. 127, no. 1, pp. 351–368, Jan. 2022. DOI: 10.1007/s11192-021-04035-5.

[28] Academia Stack Exchange, *What does authorship ordering really mean in field x?* Stack Exchange Q&A, Accessed: 2025-12-01, 2012.

[29] Samwell.ai, *What is a good h-index? practical guide for researchers*, Blog post, Accessed: 2025-12-01, Apr. 2025.

[30] Paperpile, *What is a good h-index? [with examples]*, Guide, Accessed: 2025-12-01, 2025.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[31] R. Burke, "Multisided fairness for recommendation," in *Proceedings of the ACM RecSys '17 Workshop on Responsible Recommendation*, Como, Italy: ACM, Aug. 2017, pp. 1–4. DOI: 10.1145/3109859.3109962.

[32] R. Alsaffar and S. Gauch, "Multidimensional demographic profiles for fair paper recommendation," in *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021)*, Online: SCITEPRESS, Oct. 2021, pp. 199–208. DOI: 10.5220/0010655800003064.

[33] K. Kafle, *Four trends at CVPR 2024*, Adobe Research News, Accessed: 2025-12-01, Jul. 2024.

[34] R. Islam, K. N. Keya, Z. Zeng, S. Pan, and J. Foulds, "Neural Fair Collaborative Filtering," in *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21)*, Amsterdam, Netherlands: ACM, Sep. 2021, pp. 148–159. DOI: 10.1145/3460231.3474603.

[35] H. Wu, C. Ma, B. Mitra, F. Diaz, and X. Liu, "A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation," *ACM Transactions on Information Systems (TOIS)*, vol. 41, no. 2, 47:1–47:29, 2022. DOI: 10.1145/3564285.

[36] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma, "A Survey on the Fairness of Recommender Systems," *ACM Transactions on Information Systems (TOIS)*, vol. 41, no. 3, 52:1–52:43, Feb. 2023. DOI: 10.1145/3547333.

[37] B. Bulut, B. Kaya, R. Alhajj, and M. Kaya, "A Paper Recommendation System Based on User's Research Interests," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2018, pp. 911–915. DOI: 10.1109/ASONAM.2018.8508313.

2025, © Copyright by authors, Published under agreement with IARIA - www.iaria.org