Lucy Family Institute for Data & Society

Poster Session Abstracts

THE R.I.S.E. AI CONFERENCE

October 6-8, 2025

Explore AI's potential to address societal challenges with responsible, inclusive, safe, and ethical advancements at the University of Notre Dame.

UNIVERSITY OF NOTRE DAME

Table of Contents

Tuesday, October 7, 2025

Track: Foundation Models: AI for Science Advances

#11: TOWER: Tree Organized Weighting for Evaluating Complex Instructions

Authors: Noah Ziems, Zhihan Zhang and Meng Jiang

Evaluating the ability of large language models (LLMs) to follow complex human-written instructions is essential for their deployment in real-world applications. While benchmarks like Chatbot Arena use human judges to assess model performance, they are resource-intensive and time-consuming. Alternative methods using LLMs as judges, such as AlpacaEval, MT Bench, WildBench, and InFoBench offer improvements but still do not capture that certain complex instruction aspects are more important than others to follow. To address this gap, we propose a novel evaluation metric, TOWER, that incorporates human-judged importance into the assessment of complex instruction following. We show that human annotators agree with tree-based representations of these complex instructions nearly as much as they agree with other human annotators. We release tree-based annotations of the InFoBench dataset and the corresponding evaluation code to facilitate future research.

#63: Extracting chemical reaction information using large language models

Authors: Mihir Surve, Gisela A. Gonzalez-Montiel and Olaf Wiest

Much of chemistry's collective knowledge is locked away in journal articles, patents, and electronic lab notebooks, where experimental details are reported as unstructured text. Although these records number in the millions, current structured reaction databases contain only a fraction of that information, limiting the potential of data-driven discovery. This project investigates whether large language models (LLMs) can help close this gap by extracting chemical reaction details directly from experimental text. Using a dataset of AstraZeneca electronic lab notebook entries and the Llama 3.1 Instruct model, this study tested molecule identification, role classification, and quantity extraction. The model successfully recognized reactants and solvents with high accuracy, but struggled with more ambiguous categories such as reagents, catalysts, and workup chemicals, as well as with co-references and missing product descriptions. These challenges underscore the complexity of chemical language and the need for targeted fine-tuning and annotation strategies. Despite current limitations, the results demonstrate that LLMs hold strong potential to scale up reaction curation, transforming scattered records into structured, searchable data that can accelerate synthesis planning and chemical innovation.

#82: ChefFusion: Multimodal Foundation Model Integrating Recipe and Food Image Generation

Authors: Peiyu Li, Xiaobao Huang and Nitesh Chawla

Significant work has been conducted in the domain of food computing, yet these studies typically focus on single tasks such as t2t (instruction generation from food titles and ingredients), i2t (recipe generation from food images), or t2i (food image generation from recipes). None of these approaches integrate all modalities simultaneously. To address this gap, we introduce a novel food computing foundation model that achieves true multimodality, encompassing tasks such as t2t, t2i, i2t, it2t, and t2ti. By leveraging large language models (LLMs) and pre-trained image encoder and decoder models, our model can perform a diverse array of food computing-related tasks, including food understanding, food recognition, recipe generation, and food image generation. Compared to previous models, our foundation model demonstrates a significantly broader range of capabilities and exhibits superior performance, particularly in food image generation and recipe generation tasks. We open-sourced ChefFusion at https://github.com/Peiyu-Georgia-Li/ChefFusion-Multimodal-Foundation-Model-Integrating-Recipe-and-Food-Image-Generation.git.

#84: AdSadel: Unifying Static and Dynamic Analysis with Large Language Models for Effective Mobile Ad Detection and Regulation

Authors: Shang Ma, Xusheng Xiao and Fanny Ye

Problem Statement: Mobile advertisements are extensively used in over 57% of Google Play apps as critical revenue sources. Developers integrate these ads using standard libraries (e.g., Google AdMob, Meta Ads, Applovin) or through custom-built UI widgets. However, while ads boost engagement and revenue, unregulated practices cause severe disruptions in user experiences and pose significant security risks, including intrusive ad behaviors, malware promotion, and redirection to malicious sites. Existing static analysis methods inadequately detect dynamic ad content served at runtime or custom-made ad widgets, and current dynamic approaches often fail due to inefficient UI exploration or reliance on predefined visual/textual criteria that quickly become outdated.

AI Innovation: We propose AdSadel, a novel approach harnessing Large Language Models (LLMs) to unify static ad analysis and dynamic UI exploration for effective mobile ad detection. AdSadel operates in two main phases:

Static Ad Analysis: AdSadel employs an innovative static analysis technique combining dataflow analysis and inference rules to detect diverse ad widgets through UI attributes and underlying code behaviors. It constructs a Window Transition Graph (WTG) identifying potential ad widgets and computing optimal UI exploration paths.

Ad-Oriented UI Exploration: Leveraging LLMs, AdSadel intelligently guides dynamic UI exploration. By encoding static analysis results and contextual knowledge on ad placement strategies, AdSadel provides targeted exploration guidance:

WTG-Based Guidance integrates static UI transition paths. Functionality-Based Guidance infers app functionality from metadata (name, category, description), guiding exploration towards UI widgets frequently interacted with by target users. Non-Functionality-Based Guidance leverages a knowledge base of known ad-containing UIs, utilizing similarity-based retrieval augmented generation (RAG) to prioritize exploration.

Translational Evidence: Extensive empirical evaluation demonstrates AdSadel's practical utility: testing on 156 real-world ad widgets reveals a detection accuracy of 75.86%, significantly outperforming existing methods by at least 14.84% in accuracy and achieving detection in an average of 8.52 seconds—19.37% faster. Moreover, AdSadel's deployment identifies 34 intrusive ads and uncovers 10 malware-promoting instances, underscoring its significant potential for deployment in automated ad regulation frameworks and cybersecurity policy enforcement.

RISE Dimensions: By proactively identifying intrusive advertisements and malware-promoting widgets, AdSadel safeguards user experiences and protects the integrity of mobile ecosystems. AdSadel promotes ethical mobile advertising practices through transparent detection and reporting, aiding policymakers and regulators in maintaining healthy digital environments. Additionally, the interdisciplinary nature of this work integrates insights from software security, artificial intelligence, cybersecurity policies, human-computer interaction, and ethical standards, contributing significantly towards a safer, inclusive, and responsible digital advertising ecosystem.

We provide open access to our comprehensive dataset and AdSadel's source code, enabling further research, facilitating inclusivity by allowing researchers from underserved communities to advance AI-driven cybersecurity solutions.

#85: TinyGPT-V: Efficient Multimodal Large Language Model via Small Backbones

Authors: Zhengqing Yuan, Zhaoxu Li, Weiran Huang, Lichao Sun and Fanny Ye

In recent years, multimodal large language models (MLLMs) such as GPT-4V have demonstrated remarkable advancements, excelling in a variety of vision-language tasks. Despite their prowess, the closed-source nature and computational demands of such models limit their accessibility and applicability. This study introduces TinyGPT-V, a novel open-source MLLM, designed for efficient training and inference across various vision-language tasks, including image captioning (IC) and visual question answering (VQA). Leveraging a compact yet powerful architecture, TinyGPT-V integrates the Phi-2 language model with pre-trained vision encoders, utilizing a unique mapping module for visual and linguistic information fusion. With a training regimen optimized for small backbones and employing a diverse dataset amalgam, TinyGPT-V requires significantly lower computational resources—24GB for training and as little as 8GB for inference—without compromising on performance. Our experiments demonstrate that TinyGPT-V, with its language

model 2.8 billion parameters, achieves comparable results in VQA and image inference tasks to its larger counterparts while being uniquely suited for deployment on resource-constrained devices through innovative quantization techniques. This work not only paves the way for more accessible and efficient MLLMs but also underscores the potential of smaller, optimized models in bridging the gap between high performance and computational efficiency in real-world applications. Additionally, this paper introduces a new approach to multimodal large language models using smaller backbones. Our code and training weights are available in https://github.com/DLYuanGod/TinyGPT-V.

#86: Hypergraph Representation Learning with Adaptive Broadcasting and Receiving

Authors: Tianyi Ma, Yiyue Qian, Zehong Wang, Zheyuan Zhang, Shinan Zhang, Chuxu Zhang and Fanny Ye

Hypergraphs, in contrast to general graphs, utilize hyperedges to connect multiple nodes, thereby inherently facilitating the representation of higher-order relational structures. To leverage the benefit of hypergraphs, several Hypergraph Neural Networks (HyGNNs) have been proposed to model hypergraph structures. Although existing HyGNNs achieve excellent performance to capture the complex relationships in homophilic hypergraphs, these works still face the following limitations in modeling heterophilic hypergraphs: (i) Most existing HyGNNs assume high homophily in networks, leading to limited expression ability to depict heterophilic or low homophilic hypergraph structure. (ii) A few studies attempt to address the heterophily issue through attention mechanisms to capture the importance of attribute features, which are less reliant on the homophily principle, while these attention mechanisms are ineffective for nodes in heterophilic hypergraphs. To tackle the aforementioned challenges, we propose a novel Broadcast Hyper Graph Neural Network (BHyGNN) to adaptively broadcast node information to learn more effective node representation on heterophilic hypergraphs. Specifically, we devise a novel Variational Broadcast Autoencoder Network to sample the broadcast and receive actions to propagate information between nodes and hyperedges. Moreover, we design an incorporation transformer mechanism to perform the estimated broadcast or receive actions to learn the hyperedge or node representation, incorporating the information from both sides. Extensive experiments over four benchmark heterophilic hypergraph datasets, six benchmark homophilic hypergraph datasets, and one synthetic dataset on node classification and hyperedge prediction tasks demonstrate the effectiveness of BHyGNN over baseline methods.

#115: Relevance-Aware Algorithmic Recourse

Authors: Dongwhi Kim and Nuno Moniz

As machine learning continues to gain prominence, transparency and explainability are increasingly critical. Without an understanding of these models, they can replicate and worsen human bias, adversely affecting marginalized communities. Algorithmic recourse emerges as a tool for clarifying decisions made by predictive models, providing actionable insights to alter outcomes. They answer, "What do I have to change?" to achieve the desired result. Despite their importance, current algorithmic recourse methods treat all domain values equally, which is unrealistic in real-world settings. While algorithmic recourse is extensively studied in classification tasks, its application to regression tasks remain scarce. In this paper, we propose a novel framework, Relevance-Aware Algorithmic Recourse (RAAR), that leverages the concept of relevance in applying algorithmic recourse to regression tasks. Relevance can supplement other algorithmic recourse methods, but we focus on Bayesian optimization-based methods as a baseline in this paper. We conducted multiple experiments on 15 datasets to outline how relevance influences recourses. Results demonstrate that our approach is comparable to well-known baselines while achieving greater efficiency, measured by shorter computation times and fewer iterations, and lower relative costs, indicated by more minor modifications required to achieve desired outcomes. The results, datasets, and code for replicating this study are available on GitHub.

#133: CADE: Classification with Automatic Difficulty Estimation

Authors: Ryan Cook and John Lalor

CADE: Classification with Automatic Difficulty Estimation Problem Statement Natural Language Processing (NLP) models built on transformers have improved significantly on text classification tasks in recent years. However, they often struggle to accurately measure the complexity of individual instances, which is important for ensuring data quality and model explainability. Recent work has emphasized the importance of understanding this instance-level complexity for training and evaluation of NLP methods. In this work, we propose an efficient and theoretically-grounded model for concurrently estimating the complexity of new data instances while achieving high text classification performance. We highlight the need for artificial intelligence (AI) text classifiers to be (i) Responsible in their understanding of input data and (ii) Inclusive across subgroups both in upstream data difficulty evaluations and downstream model predictions.

AI Innovation: We propose classification with automatic difficulty estimation (CADE), a semi-supervised variational autoencoder architecture to estimate both class prediction and difficulty. To develop CADE we take a multidisciplinary approach, drawing from computer science and psychometrics, specifically Item Response Theory (IRT). IRT estimates the complexity of psychometric items and has been increasingly applied in machine learning for dataset analysis. Since IRT typically requires a large dataset of responses from multiple models, we use high-dropout "burn-in" epochs to inexpensively generate outputs of various ability levels for subsequent IRT difficulty estimation. CADE reconstructs input word embeddings while making classification predictions and item difficulty estimations. Using stochastic variational inference, the multi-objective loss function minimizes distributional differences between the observed data and approximate posteriors for random variables of (i) input text features x, (ii) class label y, and (iii) response matrix R. We employ normalized bert-flow text embeddings to impose distributional assumptions that improve reconstruction of the initial text embedding x. While input text x and class labels y are observed directly from the dataset, the response matrix R is produced from the "burn-in" process and provides estimates for latent variables of model ability γ and item difficulty b. Thus, the model is able to simultaneously estimate both a probability for the class label output as well as a corresponding IRT Difficulty score for each data instance at inference. Through ablation analysis, we demonstrate that the CADE framework outperforms variational autoencoder networks of the same size, achieving similar accuracy scores to end-to-end BERT models on a psychometric classification benchmark. We also explore how performance varies across different network sizes, tasks, and text embedding methods. We validate CADE difficulty estimates via comparison to classification performance and show that the model performs worse on items with higher difficulty. Further, we examine how difficulty estimates identify both easy and hard outliers and correlate with model loss and adversarial noise. We show both (i) upstream fairness in the difficulty scores assigned to input text and (ii) downstream fairness in the output class predictions of the model across demographic subgroups.

Translational Evidence: CADE's integration with existing language models and its human-centered design affords promising deployment capabilities. Since CADE is lightweight and easily finetuned, it could be used as a calibration tool for 1 applying large language models for supervised learning tasks Given some classification task, the framework can be separately finetuned and used for inference on language model outputs. Thus, CADE provides better explainability in model calibration by identifying which model outputs are more difficult for downstream classification tasks and whether this complexity arises more from the embedding or classification process. Similar to computer-adaptive testing, CADE might also be used to analyze and refine data quality. By identifying difficult examples, CADE could be used to help the language model generate outputs that are more or less "adversarial" to downstream processors, whether human or machine. Since the framework takes a human-centered design through IRT, the difficulty scores generated by CADE are indicative of how much a diverse subject population will struggle to understand the model's text output.

RISE Dimension: Improving the model's understanding of data complexity and quality emphasizes the development of responsible AI. We measure this instance-level complexity by an established, theoretically informed method that is directly related to the expected classification performance. Thus, we shine light into black box models by providing downstream users not only with output predictions but also difficulty scores to proxy the model's understanding of the text. By highlighting examples that models find difficult to understand, we enable more informed decision-making concerning model performance on tasks such as in-context learning, leaderboard construction, and ambiguity resolution from limited text. Since instance complexity deals with the likelihood of misclassification, it is intrinsically linked with

fairness considerations. Our difficulty estimates come from reconstructing a response matrix provided by (mis)classifications from non-human model subjects. Thus, these artificial crowds may avoid traditional human biases by being trained on more inclusive training text (than individual human experience) but risk central tendencies of statistical learning that favor the majority class. However, we demonstrate that difficulty scores are generally similarly distributed across protected subgroups across protected subgroups such that our estimates provide a more general type of linguistic difficulty that applies to individuals across demographics. CADE is able to estimate the difficulty of unseen data instances with high predictive performance. Aside from its predictive value, our model provides insights into machine understanding of text complexity as an auxiliary task to its classification. It incorporates mechanisms for assessing data quality and fairness by design, contributing to responsible and inclusive development of AI.

#136: Embracing Missingness: Enhancing Similarity Measures via Probabilistic Embeddings

Authors: Jerrick Gerald, Nitesh Chawla and Keith Feldman

Missing data is a pervasive challenge in real-world tabular datasets, often leading traditional representation learning models to rely on imputation techniques that ignore uncertainty and distort similarity estimates. We introduce Variational Uncertainty Embedding (VUE), a self-supervised probabilistic framework that explicitly models missingness as an informative signal rather than a nuisance. VUE learns distributions over embeddings by incorporating both observed features and a binary missingness mask, allowing it to represent uncertainty directly within the embedding space. Our contrastive training objective leverages mean and variance representations, regularized with KL divergence, and includes a missingness-aware penalty to discourage overconfidence in partially observed data. In extensive experiments, VUE consistently outperforms existing methods such as SCARF in preserving local data structure under uncertainty, as measured by trustworthiness scores. Beyond technical improvements, VUE offers meaningful translational impact—particularly in safety-critical domains such as healthcare and social sciences—where treating missingness as uncertainty can lead to more equitable, transparent, and cautious decision-making. This work aligns with RISE principles of responsibility and safety by addressing structural biases that emerge when missing data is overlooked or improperly handled.

#156: Dimension Reduction via Data Integration

Authors: Bingxue An and Tiffany Tang

A plethora of dimension reduction methods have been developed to visualize high-dimensional data in low dimensions. However, different dimension reduction methods often output different visualizations, and there are many challenges that make it difficult for researchers to determine which visualization is best. We thus propose a novel consensus dimension reduction framework, which summarizes multiple visualizations into a single “consensus” visualization. Here, we leverage ideas from data integration in order to identify the patterns that are most stable or shared across the many different dimension reduction visualizations and subsequently visualize this shared structure in a single low-dimensional plot. We demonstrate that this consensus visualization effectively identifies and preserves the shared low-dimensional data structure through extensive simulations and real-world case studies.

We further highlight our method's robustness to the choice of dimension reduction method and/or hyperparameters – a highly-desirable property when working towards trustworthy and reproducible data science. This consensus visualization framework contributes to the responsibility dimension. It provides researchers with an effective way to capture the intrinsic structure of datasets and minimize the risk of misinterpretation. Dimension reduction methods are widely used in many fields, but the results can be misleading if they are not carefully evaluated or the parameters are poorly tuned. By selecting a set of candidate dimension reduction methods based on exploratory data analysis, researchers can apply our work to generate a consensus visualization that is both reliable and informative. This technique is highly applicable across a wide range of fields. In healthcare, for instance, analyzing genomic datasets with thousands of genes presents challenges.

Applying this framework allows us to effectively summarize insights from multiple dimension reduction methods, making it easier to identify patterns associated with diseases or medical responses while preserving as much relevant information as possible. Furthermore, the application of this project is not limited to dimension reduction and data visualization. It can also be applied to other machine learning problems, such as clustering, classification, and prediction. We can combine results from different clustering algorithms and provide a consensus clustering to improve the robustness and accuracy of the analysis.

#170: Modeling Edge-Rich Graphs using Neural Networks

Authors: Shaochun Li, John Lalor and Ahmed Abbasi

Traditional graph-based methodologies in spatio-temporal analysis largely adopt node-centric frameworks, leveraging rich node information (e.g., object attributes, mobility patterns) to model spatial-temporal data. However, edge-aware graph modeling with rich edge information remains underexplored, despite edges naturally capturing object interaction patterns. This gap limits the ability to model complex spatio-temporal phenomena such as object movement forecasting. Identifying object trajectories in spatio-temporal data is critical for real-world forecasting applications, including personalized marketing, traffic routing, and event detection. However, predicting object movements across time and space is non-trivial due to the large number of possible prediction outcomes and challenges in the potential feature space stemming from dynamic and nuanced movement patterns that require careful consideration for spatial granularity, temporal resolution, and object edge heterogeneity.

To address these challenges, we propose an edge-conditioned convolutional graph neural network based approach that models spatio-temporal data as context-enriched graphs for enhanced object trajectory forecasting. Our model learns latent trajectory patterns from the underlying graphs by projecting nodes and dynamic edges into continuous embedding spaces that can be utilized by statistical models. More specifically, the proposed method employs a spatial cell graph formulation mechanism to manage spatio-temporal resolution in conjunction with edge-conditioning node embeddings and graph embeddings to account for object edge heterogeneity and macro-level movement patterns, respectively. The embeddings also encompass structural, temporal, and spatial properties of the graphs and can be visualized for better interpretation. We evaluate the proposed method against state-of-the-art feature-based and graph neural network methods for movement prediction on three real-world data sets. The experimental results suggest that our proposed method is significantly more accurate at predicting object movements, with multi-class ROC and accuracy values that are at least 3 to 4 percentage points higher than benchmark comparison methods. These results are consistent across a wide range of parameter settings. Further, ablation analysis shows the effectiveness of the key underlying components of our proposed method (including the edge-conditioned node and graph embeddings). Moreover, comprehensive simulation experiments under varying data structures demonstrate the robustness of the proposed model. With increases in the collection, storage, and usage of big data, as well as a greater appetite for agility and responsiveness in near real-time environments, spatio-temporal modeling will continue to play a prominent role. How to effectively represent spatial granularity, temporal resolution, and object heterogeneity manifesting in graph edges as opposed to graph nodes are critical considerations for models that can successfully predict movements across time and space. Our work has important downstream implications for real-time predictive modeling and decision-making in contexts such as health, security, smart cities, and marketing. Our research aligns with the safety dimension of RISE AI, and demonstrates significant potential for real-world deployment in traffic control and pedestrian trajectory forecasting, bridging AI innovation with smart city demands.

Track: Reimagining Global Governance and Policymaking

#5: Constructing Political Indices using Machine Learning: The Case of Official Patriotism

Authors: Peitong Jing

Machine Learning (ML) models enable scientists and policy makers to measure complex social phenomena such as democracy and nationalism using large observations. While they significantly reduce the cost spent on processing data, many ML models lack transparency. Consequently, policy practitioners and even scientists themselves cannot interpret what elements contribute to measuring the final concept. Focusing on state-promoted nationalism as a latent concept, this paper explores black-box and white-box Machine Learning models for indexing the phenomenon over time for countries in the world. When an exogenously determined proxy indicator is available, we can use this indicator and ML to measure latent concepts to explain complex socio-political phenomenon. In this case, I find a proxy indicator for official nationalism and its related indicators in high-quality datasets based on individual and expert surveys such as Varieties of Democracy and Variety of Indoctrination. One major challenge to utilize such large datasets lies in selecting important variables or indicators. ML models can solve this challenge, but users and designers need to be mindful of model interpretability. I argue that model transparency is an important evaluation criterion along with predictive power. Specifically, I propose a procedure that involves four steps: (1) identifying a proxy indicator for the latent concept from existing datasets, (2) applying Machine Learning tools to select relevant indicators, (3) determining the factors and components that make up the concept, and (4) employing Bayesian latent variable modeling to estimate the latent concept. I demonstrate this approach by building an index for official nationalism across the world. The index opens up avenues of comparative research on official nationalism and nationalistic propaganda. Using China and India as two country cases, I further demonstrate how the index and qualitive case studies mutually inform each other to deepen our understanding of official nationalism.

#14: Transforming Forensic Text Analysis with Large Language Models: A Case for Scalable and Insightful Justice

Authors: Filipe da Luz Lemos, Rubens de Faria and Thiago dos Santos Cavali

Problem Statement: Modern forensic investigations often result in the seizure of vast amounts of textual data—such as cellphone messages, social media conversations, and call transcripts. While current forensic workflows focus primarily on the extraction and validation of such data, the sheer volume of information can overwhelm defendants, public defenders, prosecutors, and judges alike. Traditional analysis methods—whether manual or keyword-based—struggle to make sense of these datasets in a systematic way, compromising fairness and efficiency, particularly for under-resourced parties.

AI Innovation: We propose the use of Large Language Models (LLMs) to enable a scientific, scalable post-extraction analysis of forensic data. Without requiring further fine-tuning, LLMs can immediately assist in:

- Robust entity recognition, capturing spelling variants and nicknames (e.g., "Filipe" vs. "Felipe", "Doctor" as a nickname);
- Sentiment and intention analysis, detecting implicit threats, cooperation, or behavioral patterns
- Automatic summarization of large volumes of communication;
- Contextual association and relationship mapping across messages and transcripts.

By analyzing linguistic context and associating entities across multiple conversations, LLMs help structure unorganized datasets into interpretable evidence, supporting a scientific approach to forensic interpretation. Outputs are designed to be interpretable and are subject to mandatory validation by forensic professionals to ensure accuracy and prevent hallucinations or misinterpretations. Translational Evidence A practical illustration occurred during Brazil's Vaza Jato investigation, where defense teams were initially provided with raw, unorganized data extracted from a hacker's devices. Through systematic data cleaning and the application of LLMs, it became possible to detect critical communications—including exculpatory evidence—that would have been virtually inaccessible otherwise due to the dataset's size and the

heavy use of aliases among participants. This demonstrates how LLMs can not only reduce review times from months to days but also elevate the scientific rigor of forensic text analysis, making advanced forensic capabilities accessible to all defendants, regardless of resources.

RISE Dimensions Alignment

- Inclusion: Expands access to forensic analysis for underserved populations through free or low-cost LLM-based tools.
- Responsibility: Embeds human oversight into the analysis process, with transparent and interpretable outputs.
- Ethics: Protects individual rights by enforcing data anonymization and compliance with privacy regulations (LGPD, GDPR) and by auditing for bias.
- Safety: Strengthens judicial integrity through human validation, operational guidelines, and structured error mitigation.

#45: Expanding Legal Access

Authors: Susan Azyndar and Thomas Mills

Problem Statement: Many people cannot afford a lawyer but do not qualify for legal aid, leaving them to navigate the legal system on their own. Without reliable legal information, these individuals struggle to understand court procedures, complete legal forms correctly, or assert their rights effectively. Traditional legal resources, often written in complex legalese, can be difficult to interpret and require legal training to leverage successfully. Self-help options vary widely in both quality and accessibility. AI holds promise for developing remedies to fill this gap by making legal information more accessible. Simply entering a prompt into ChatGPT or one of its competitors, though, is not an adequate solution. Prompt engineering in these chatbots requires legal knowledge in order to receive relevant results, and even then, the basis of the answer, the underlying legal authority, isn't always clear. Moreover, answers from these chatbots are not vetted, as the continuing spate of cases involving lawyers who cite fake cases attests. There remains a critical need for accurate, jurisdiction-specific legal information for self-represented litigants and underserved communities.

AI Innovation: Drawing on our work as law librarians, we will discuss a new way to address the justice gap: an AI-driven legal information tool designed to help bridge this justice gap. Unlike generic AI chatbots, this product prioritizes reliable, plain-language legal guidance drawn from verified sources. We will discuss key considerations, such as ensuring accuracy, addressing bias, and maintaining ethical standards in AI-generated legal content. It will also highlight how libraries, courts, and legal aid organizations can leverage AI responsibly to support public legal education.

Translation Evidence: Cost is a significant barrier to access to legal services. According to the Legal Services Corporation, "Among Americans who did not seek legal assistance for a civil matter they experienced over the past three years, nearly a third (32%) did not do so because they were worried about the cost. A freely available AI system would offer many more people access to a sounder approach to self-help. This system would counsel users to seek legal advice from a practitioner as well, to adhere to the ethical requirements around the unauthorized practice of law.

RISE Dimensions: This proposal focuses on the inclusion dimension, helping those who currently cannot afford a lawyer to access quality legal information freely. Format: We prefer a poster and are open to a panel or fireside chat.

#62: Investigating the Influence of Credibility Indicators and Social Engagement on News Perception

Authors: Adnan Hoq, Matthew Facciani and Tim Weninger

The rapid proliferation of manipulated content and polarization in contemporary digital media presents a critical socio-technical challenge, severely impacting democratic discourse, public trust, and societal cohesion (Shen et al., 2019; Wang et al., 2022). Social media platforms, driven by engagement metrics (likes, shares, comments), algorithmically shape information exposure and credibility judgments (Hermida et al., 2012; Muchnik et al., 2013, Messing & Westwood, 2014). Concurrently, novel algorithmic credibility indicators, such as AI-generated content assessments and institutional fact-checking labels, are increasingly integrated into these platforms, but their interplay with social cues

and political identity remains largely unexplored (Glenski & Weninger, 2017; Scheufele & Tewksbury, 2007). This lack of understanding hinders the development of effective platform policies and educational interventions, disproportionately affecting underserved communities, where manipulated content can exacerbate existing societal inequalities. To address these challenges, this study introduces a novel interdisciplinary methodological approach combining experimental social science with applied artificial intelligence techniques. Specifically, we designed a rigorous, mixed-design online experiment involving 1,000 diverse participants recruited through Prolific, systematically evaluating how AI-generated credibility assessments (via ChatGPT-generated feedback), institutional credibility indicators (via GroundNews media bias breakdowns), and social engagement metrics jointly influence perceptions of news accuracy and shareability. Participants rate a carefully curated set of 21 news articles, varied explicitly across political alignment (ingroup/outgroup), social engagement levels (none, low, medium, high), and feedback conditions (Control, GroundNews Breakdown, GroundNews Inverted, and ChatGPT feedback). The methodological innovation lies in employing generative AI (ChatGPT) as an active intervention to simulate scalable, real-time credibility assessments alongside traditional institutional signals, enabling robust empirical comparisons and revealing nuanced interactions between algorithmic assessments, identity-driven biases, and social proof effects. The AI-driven credibility assessment methodology developed in this research holds substantial promise for translational deployment, significantly informing platform design, algorithmic moderation policies, and digital literacy strategies. Empirical findings from the experiment will provide actionable insights for social media companies aiming to design more effective credibility interventions, optimize content moderation, and counteract manipulated content. Moreover, by delineating the conditions under which political identity biases override algorithmic credibility signals, policymakers can better target and customize public education campaigns to mitigate polarization. Given the scalable nature of AI-driven interventions, such approaches could feasibly be integrated directly into social media ecosystems, extending reach to diverse user groups, including underserved communities who are disproportionately vulnerable to manipulated content and polarization due to limited media literacy resources and heightened reliance on algorithmically mediated content. This translational potential underscores the significant socio-economic impacts achievable through evidence-based integration of AI and media credibility indicators within digital news environments. This research directly engages with multiple RISE dimensions, emphasizing responsibility, inclusion, safety, and ethics in AI innovation and deployment. Firstly, responsibility is addressed through transparent, empirically grounded assessments of AI-generated credibility indicators, explicitly evaluating potential biases and limitations. Secondly, inclusion is central to the study's design; by recruiting a demographically diverse participant pool (across race, gender, geographic location, education, and political affiliation), the research ensures representation of voices traditionally underrepresented in algorithmic impact studies. Understanding differential effects across these groups further facilitates inclusive platform design and equitable access to reliable information. Thirdly, safety is explicitly considered, as the study provides foundational knowledge to mitigate manipulated content's harmful consequences—such as public health manipulation or political manipulation—enhancing societal resilience and collective decision-making capacity. Lastly, ethics is foregrounded through critical examination of how algorithmic credibility indicators interact with political biases and herd behavior, proactively considering potential ethical dilemmas related to algorithmic transparency, trustworthiness, autonomy, and manipulation. By centering these RISE dimensions, the proposed AI approach not only innovates technologically but also advances socially responsible, inclusive, and ethical information systems design. Ultimately, this research embodies an interdisciplinary, human-centered approach to AI, leveraging advanced computational techniques alongside rigorous social science methodologies to foster informed, resilient, and inclusive digital societies.

#65: Balancing between Efficiency and Inefficiency: When Economic Sanctions Meet AI/ML

Authors: Sanghyun Han

This article examines the conditions under which the US government gains or loses efficiency in federal affairs, particularly national security, through the adoption of artificial intelligence and machine-learning (AI/ML) technologies. Tracing the evolving approaches of US administrations, it highlights a shift from exploratory applications of AI/ML to more adoption in federal operations, including high-stakes areas such as national security. Using the US system of economic sanctions as a case study, it explores a paradoxical dynamic: while AI/ML has the potential to enhance efficiency by improving tasks such as licensing and tracking illicit financial transactions, its implementation introduces mediating factors that temper these gains. While tasks such as licensing and tracking illicit financial transactions can benefit from AI/ML adoption, inefficiencies arise when these systems fail to deliver accurate intelligence or decisions, require extensive explanations and additional validation, or impose high infrastructure costs, including investments in human capital and protective measures. It aims to alleviate uncertainty regarding AI/ML applications in national security and initiate broader discussions on their role in federal affairs, providing practical insights to guide future deliberations.

#67: Quantifying Responsibility for Cross-state Air Pollution: An airshed approach

Authors: Wenxu Liao, Richard Marcantonio and Paola Crippa

Air pollution is a leading environmental health risk, responsible for significant premature mortality worldwide. Due to its transboundary nature, pollution management poses complex regulatory challenges, especially in attributing responsibility for cross-state emissions. Traditional approaches often fail to adequately quantify pollution flows and their socio-economic impacts, necessitating an advanced methodological framework for equitable policy interventions.

This study presents an AI-driven airshed modeling approach designed to systematically attribute responsibility for fine particulate matter (PM2.5) pollution across U.S. states. Our framework applies unsupervised machine learning and clustering techniques to dynamically define airsheds, which are geographical regions shaped by atmospheric transport patterns. This approach enables the identification of pollution sources and the evaluation of their impact on affected populations. Unlike chemical transport models, which demand extensive computational resources, our AI-based method integrates real-time meteorological data with high-resolution pollution measurements, offering a scalable and dynamic assessment of cross-state air quality.

Our findings reveal that cross-state pollution contributes to approximately 40% of premature deaths linked to PM2.5 exposure. Some states emerge as net exporters of pollution, while others bear a disproportionate share of the health burden. By quantifying these disparities, our approach provides a robust foundation for policymakers to design fair and effective pollution control strategies, ensuring accountability and optimizing air quality management nationwide. Beyond regulatory applications, our framework enhances evidence-based decision-making by enabling precise attribution of emissions. It also underscores disparities in pollution exposure, emphasizing the urgency of protective measures for vulnerable communities disproportionately affected by poor air quality. Addressing cross-state pollution is not just a matter of regulatory efficiency. It is a step toward environmental justice, ensuring that high-emission states take responsibility for the broader health impacts of their pollution.

By advancing our scientific understanding of transboundary pollution and enabling targeted policy actions, this data-driven airshed modeling approach represents a transformative tool in air quality governance. It supports a future where environmental management is not only data-driven and effective but also equitable and inclusive, paving the way for responsible and ethical AI applications in public health and environmental policy.

#70: Building KGs from the Colombian Truth Commission's Archives

Authors: Anna Sokol, Leonardo Andres Ibañez Tirado, Viviana Gomez Leon, Maria Paula Prada Ramírez, Luis Gabriel Moreno Sandoval, Matthew L. Sisk, Josefina Echavarria Alvarez and Nitesh V. Chawla

Problem Statement: The Colombian Truth Commission has collected thousands of anonymous testimonies about human rights violations during Colombia's long armed conflict. These testimonies come from survivors and witnesses who described violence, forced displacements, and other terrible acts. Because these testimonies are unstructured and anonymous, it's hard to see connections between events, places, and people involved. Without a good way to analyze these testimonies, researchers, policymakers, and human rights advocates cannot easily identify overlapping stories or verify supporting details. This makes it difficult to accurately reconstruct history, allocate resources effectively, or develop reconciliation programs for affected communities.

Using advanced AI techniques could transform this scattered data into an organized, searchable format that respects confidentiality while revealing important social and political patterns. The main challenge is converting these anonymous testimonies into a structured format that keeps witnesses' identities confidential while allowing for deep analysis across multiple documents. Simple keyword searches don't work well with these complex narratives about human rights abuses. For example, testimonies often use placeholders like “Witness 123” instead of real names, or vague terms like “the local official” to protect identities. While necessary for privacy, this makes it very difficult to find patterns or confirm connections between different accounts using regular search methods. The problem becomes even more complex when different testimonies describe similar incidents using different words or time references. Our research question is: How can we use AI methods to extract, connect, and represent the hidden structures in these narratives, while maintaining ethical standards of anonymity and still uncovering valuable historical insights? By addressing this challenge, we aim to balance the need for rigorous analysis with the need to protect the dignity and privacy of trauma survivors.

AI Innovation: To address this challenge, we propose using cutting-edge natural language processing techniques to convert unstructured, anonymous narratives into a dynamic Knowledge Graph. Our framework uses transformer-based large language models (LLMs) with prompt-driven extraction methods to identify key entities, events, and their relationships -even when personal identifiers have been deliberately hidden. For example, the system recognizes event triggers like “massacre,” “kidnapped,” or “forced displacement” by analyzing context clues such as time markers and geographic indicators, building event nodes that connect various testimonies.

By integrating KG search and traversal, the framework enables advanced queries, and recommendation functions, moving beyond simple document retrieval to interactive, knowledge-driven exploration. This approach not only helps the users better understand historical events but also serves as a model for using AI in other contexts where data anonymity and sensitivity are important. In practice, the KG can be leveraged to power recommendation systems for chatbots, enabling them to extract semantic relationships between entities and deliver contextually relevant suggestions based on user interactions and historical patterns. When users research specific events, the chatbot can leverage the KG to identify broader patterns across testimonies, suggesting connections like “Similar displacement events occurred in neighboring municipalities within the same month, suggesting a coordinated campaign” without ever compromising witness identities.

Translational Evidence: This AI solution has significant real-world applications beyond academic research, with important socio-economic and policy implications. By organizing anonymous conflict testimonies into a searchable KG, the framework gives policymakers, legal investigators, and human rights organizations a powerful tool to identify patterns of violence, recurring incidents, and changes in armed group behavior over time. For instance, if multiple interviews show a pattern of forced displacement in a specific region over several years, these insights can directly inform targeted reparation policies, resource allocation, and legal investigations. Similar projects—such as automated KG construction for historical documents from the Chilean dictatorship—demonstrate that this approach can effectively support transitional justice and build public trust through transparent, data-driven insights. The technology's ability to uncover hidden connections among separate testimonies can transform historical analysis and policy-making, ensuring that even anonymous data can contribute to a better understanding of past conflicts and inform future reconciliation efforts.

RISE Dimensions: Responsibility: We use controlled entity linking to protect vulnerable witnesses' identities while still allowing for rigorous academic and policy analysis. Inclusion: The framework handles multilingual testimonies and diverse narrative structures, ensuring fair representation of voices from different regions and backgrounds in the KG. Safety: We implement robust measures to prevent re-identification, without ever attempting to reveal hidden identities, including selective external linking and uncertainty indicators for ambiguous data, reducing risks associated with handling sensitive information." Ethics: Transparency in the extraction process, combined with explainable AI methods, allows human reviewers to audit and verify each extracted fact, ensuring the system maintains ethical standards when processing data from conflict settings.

#122: The Algorithmic Muse: A Human-AI Duet in Creative Verse and Vision: An Autoethnographic Exploration of Accessible and Inclusive Artistic Expression Through a Human-AI Dyad, Viewed Through a RISE Lens

Authors: Mechelle Gilford

This paper addresses the grand challenge of fostering accessible and inclusive artistic expression in an increasingly digital world, particularly for underserved communities who may face barriers to traditional art creation and engagement. The AI innovation presented is a methodological framework for human-AI co-creation using large language models (LLMs) specifically Google's Gemini as a collaborative partner for artists and poets, leveraging the AI's capacity for novel textual and visual generation based on whimsical prompts inspired by surrealism and glitch aesthetics. Translational evidence is demonstrated through the autoethnographic case study of a Glitch Artist and Poet collaborating with Gemini to produce unique poems and digital artworks. This approach holds deployment potential in educational settings (innovative instruction in art),therapeutic contexts (AI-assisted art therapy),and community arts programs, offering new avenues for creative participation and self-expression that can empower individuals regardless of physical or technical limitations, thus contributing to socio-economic well-being through enhanced creative agency and social inclusion.

This collaboration aligns with the RISE dimensions by:

1. promoting Responsibility through intentional human guidance of the AI's creative output;
2. fostering Inclusion by exploring AI's capacity to create accessible art and facilitate creative expression for diverse individuals, including those from underserved communities;
3. implicitly considering Safety by focusing on positive and expressive creative outcomes; and
4. encouraging Ethics through reflexive examination of the evolving human-AI relationship in creative endeavors. This interdisciplinary approach integrates computer science (AI), humanities (art, poetry, surrealism), social sciences (accessibility, inclusion), and education (art therapy, instruction), potentially amplifying diverse voices and focusing on underserved communities by offering novel,technologically mediated pathways to artistic creation and engagement.

#127: Using Artificial Intelligence to Prevent Wrongful Convictions: A Framework for Responsible and Ethical Justice Systems

Authors: Madiha Mirza

Problem Statement: The justice system, designed to uphold fairness and protect human rights, has long struggled with the devastating reality of wrongful convictions. Factors such as racial bias, flawed forensic practices, eyewitness misidentification, and systemic inequalities disproportionately harm marginalized communities, eroding trust in institutions meant to safeguard them. Despite growing awareness, tools to systematically prevent wrongful convictions remain limited. This solution explores how artificial intelligence, when developed responsibly, can serve as a safeguard against wrongful convictions, supporting a justice system that truly embodies dignity, fairness, and accountability for all.

AI Innovation: This solution proposes a responsible AI framework for wrongful conviction prevention that integrates: ▪ Bias Detection Algorithms: Identifying systemic patterns in case outcomes, evidence interpretation, and sentencing disparities. ▪ Case Review Support Systems: Using AI to prioritize potential wrongful conviction cases by flagging

inconsistencies, unreliable witness testimonies, or forensic anomalies. ▪ Transparent Evidence Evaluation Models: Assisting legal professionals by providing explainable risk assessments based on cross-validated forensic and procedural indicators. Unlike traditional predictive policing or risk scoring systems, which can exacerbate biases, this framework is explicitly designed to reduce systemic harm and operate under strict ethical guidelines emphasizing transparency, accountability, and human oversight.

Translational Evidence: Early-stage projects, such as the Innocence Project's use of forensic data audits and machine learning applications in legal analytics, demonstrate AI's potential in identifying wrongful convictions. Pilot initiatives integrating natural language processing with case law review have successfully flagged problematic patterns, e.g., unreliable expert testimony clusters, that correlate with later overturned convictions. Building on these precedents, this solution outlines how a rigorously ethical, human-centered AI system could be piloted in partnership with public defenders, innocence organizations, and judicial oversight bodies to proactively protect vulnerable populations.

RISE Dimensions:

- Responsibility: AI systems are designed to support, not replace, human judgment, with clear audit trails and accountability mechanisms ensuring transparency and contestability.
- Inclusion: Development centers on underserved communities disproportionately affected by wrongful convictions, embedding racial, socioeconomic, and cultural sensitivity into system design.
- Safety: The system's primary function is harm prevention, identifying risks of injustice before they result in irreversible outcomes.
- Ethics: Ethical guidelines are built into every stage, from data sourcing to model deployment, prioritizing human dignity, fairness, and the right to due process.

#137: Results of the Oxford Collaboration on Theology and Artificial Intelligence

Authors: Lyndon Drake

This paper presents results from the Oxford Collaboration on Theology and Artificial Intelligence, a multidisciplinary AI ethics project with a substantive majority world focus. The project has brought together a highly diverse network of theologians and AI experts, with the goal of developing an ethos and a professional commitment for those working in AI. The project has developed novel contributions to theologically-informed AI ethics in methodology, topics addressed, and substantive theological ethics. The ethos and professional commitment specifically address the dimension of ethics. The project has four major components: listening to practitioners in science, engineering, and industry; building a network of practitioners and theologians who are willing to collaborate; producing theologically-informed ethical material; and proposing the Oxford Oath, a professional commitment for AI practitioners.

AI ethics often focuses on large-scale issues, such as the nature of intelligence or human personhood, consciousness, and societal risks arising from Artificial General Intelligence. The motivating question for this project arose from AI engineers: what could theologians say to address the ethical challenges faced by AI practitioners? A problem commonly faced by those involved in advancing the science and engineering of AI systems, and building companies that develop and use AI systems, is that their vocational realities present different and more granular ethical challenges from those large-scale questions. Alongside that vocational issue, much AI ethics in the public sphere is explicitly secular, and devout or religiously-sensitive practitioners have approached us seeking AI ethics that is theologically-informed.

The project addresses that problem in a methodologically-novel manner, by giving significant attention to contexts where theologians are invited to listen to practitioners. These include scientists introducing the technical features of their discipline, especially the breadth of AI science — in other words, beyond chatbots and LLMs. Engineers have described the vocational practice of their daily work, rather than just showing the outputs of their systems. Corporate leaders have explained how capital allocation for AI, which has immense social importance when AI companies function on a scale comparable to some countries, works in practice.

These vocational realities are also differentiated across the world, with those in the majority world posing additional and sometimes different ethical questions. While the world as a whole has a growing religious majority, this is especially true

outside the cultural West. Many of those involved in the development and use of AI wish to have theologically-informed ethical contributions to their vocational realities.

As well as surfacing these granular, vocational issues and questions, a second major methodological contribution is the network of theologians who are engaging in developing new theological resources. At one level, it is important for AI ethics to function persuasively in a religious world, given that many societies and a majority of humans are religious adherents, so for ethical discourse to affect the world, it needs to address theological concerns and interact with theological reasoning. Additionally, many of the ethical questions which arise in the context of AI (for example, of what it means to be human) are ones which theologians have a long history of addressing. Many of the theologians in the network are both scientists and theologians, and so the project also demonstrates the value of subject expertise in both science and theology, seeking to avoid a naïve reactionary posture towards the scientific and social developments around AI.

Among the central concerns which the project addresses are: medical uses of AI, which offer increases in access to expertise and pose challenges of equity and bias; issues of language, especially in relation to minoritised languages; the formative impact of interaction with AI systems on the human person; the export of dehumanising labour to the poor and to the majority world; and the opportunities arising from AI systems to contribute back to theology as a discipline. In each case, the project addresses the more granular concerns which arise from a careful attention to work in AI, and to the scientific distinctions between different types of AI and the technical details of AI engineering. The outcomes of the project are not intended to serve as a final word. Instead, the ethos and professional commitment we are producing offer an example of methodology — deep engagement with practitioners and the majority world, and with theology — and provide a network which can continue to reflect and develop in the context of the profound social change that ongoing development and deployment of AI systems will produce in the years to come.

#150: Automatic Identification of Global Moderators with Second-order SHAP Interaction Values

Authors: Deng Pan and Jing Peng

In domains such as social sciences, medicine, public health, and education, moderation effects in regression analyses are traditionally hypothesized based on prior theoretical frameworks and validated through models incorporating manually selected interaction terms. However, this manual and theory-driven approach relies heavily on domain expertise, limiting scalability and potentially overlooking important moderators. To address these limitations, we propose an automated, data-driven framework for identifying global moderation effects using second-order SHAP interaction values. Our approach trains tree-based models on the data and employs TreeSHAP to compute interaction values between feature pairs. For each candidate moderator, we aggregate these interaction values across instances, evaluate their directional consistency, and statistically assess their significance. Through experiments on both synthetic and real-world datasets (such as the Resilient Community Database), we will demonstrate that our framework not only recovers known moderators but also uncovers novel interaction patterns. By delivering interpretable and scalable insights into moderation effects, this method supports responsible, data-informed decision-making across disciplines.

Introduction: Mediation and moderation are foundational tools in regression analysis, enabling researchers to understand how a third variable influences the relationship between a predictor and an outcome. Mediation seeks to explain the underlying mechanism, while moderation examines how the strength or direction of the relationship varies across subgroups. Classic examples of moderation include how income influences the effect of education on political participation, or how gender moderates perceptions of public policy effectiveness. Traditionally, moderation effects are hypothesized from prior theoretical frameworks and validated by fitting regression models with manually selected interaction terms. While this approach benefits from interpretability, it is limited by its reliance on domain expertise and manual specification, which hinders scalability and may miss subtle or unexpected moderators. To overcome these limitations, we propose an interpretable machine learning framework that automatically identifies global moderation effects using second-order SHAP interaction values.

Method: We introduce a three-stage framework to identify global moderators in complex models using second-order SHAP interaction values:

- Model Training with Tree-Based Methods We fit interpretable, high-capacity tree-based models (such as XGBoost or LightGBM) on domain-specific datasets. These models are well-suited for capturing non-linear relationships and interactions and support exact SHAP value computation via TreeSHAP.
- Extraction of Second-Order SHAP Interaction Values We use TreeSHAP to compute second-order SHAP interaction values for all feature pairs. For a given pair of features (i, j), the interaction value quantifies how the influence of feature i on the prediction depends on feature j.
- Aggregation and Statistical Evaluation of Moderator Effects. For each candidate moderator j with respect to feature i, we:
    - Compute the mean interaction value to estimate the average strength and direction of the moderation effect.
    - Assess consistency across instances by measuring polarity: the proportion of instances where the interaction has a consistent sign.
    - Conduct statistical hypothesis testing (e.g., one-sample t-tests) to determine significance.
- A feature j is deemed a global moderator of i if it consistently and significantly alters i's effect on the prediction. Results will be visualized using interaction heatmaps, stratified dependence plots, and ranked moderator lists.
- Empirical Validation:We validate the proposed framework on both synthetic datasets (with known ground-truth moderators) and real-world datasets such as the Resilient Community Database curated by our team. Effectiveness is assessed by the ability to:
    - Recover known moderators,
    - Discover novel, plausible interactions, and
    - Provide interpretable explanations.

#159: Location, Location, Location: Viewing Generative AI's Latent Space Through the Lens of Property Law

Authors: Colton Crum, Patrick Flynn and Adam Czajka

Generative artificial intelligence (AI) is arguably the most prevalent and sought-after application of contemporary AI. These classes of AI models are capable of generating synthetic content that is indistinguishable from reality, blurring the lines between fact and fiction. Furthermore, these models act at the whim of a user-based prompt, obedient in generating content that is only bounded by the imagination of the user. Since AI-generated content is increasingly being compared to both inventions and discoveries. As generative models begin to generate discoveries, we must ask, "Who owns those discoveries?" and "Where do those discoveries reside?" The crucial element often missed within these discussions resides not within the user, but rather within the labyrinth which powers any generative AI model – its latent space. In this paper, we present the idea of "latent space as a service", in which cryptography-based "fencing" of the latent sub-manifolds regulate the usage of that space depending on the application and target audience. First, we describe a model's latent space, which is an embedding landscape that encodes the content provided during training and allows generative models to generate content. Within this space, a manifold (or terrain) exists where all of the model's potential creations reside. Second, we connect this latent space manifold with user-based inputs, which are transformed to reside in a given location within the latent space, from which they can drive the generation of model output. By elucidating the relationship between user inputs and a generative model's latent space, a newfound view of generative AI begins to emerge. Generative AI and its discoveries may be the digital equivalent of physical territory, property, or even real estate. Furthermore, this analogy is supported by ongoing AI governance strategies by implementing "guardrails" to prevent models from traveling "off-path. Additionally, efforts are made to prevent users from obtaining outputs within a model's latent space, which are frequently circumnavigated by "jailbreaking" techniques. Both strategies create explicit boundaries or otherwise locations from which a user is denied access to a particular location within the model's latent space. Under this newfound lens, generative AI and its uses raise several legal considerations related to digital property laws, intellectual property, and potential hazards that come from traversing through AI's enigmatic terrain. Finally, the revitalized perspective forges a novel path forward for Generative AI, creating unique applications within privacy rights, data licensing, and cryptography within the largely untapped digital landscape that awaits us.

#164: Measuring and Mitigating Homelessness Bias: Leveraging AI for Social Impact

Authors: Jonathan Karr, Ben Herbst, Matthew Hauenstein, Georgina Curto and Nitesh Chawla

1. Problem Statement: Bias towards people experiencing homelessness (PEH) is prevalent in online spaces. We leverage natural language processing (NLP) and large language models (LLMs) to identify, classify, and measure bias using geolocalized data collected from X (formerly Twitter), Reddit, meeting minutes, and news articles across the United States. The results of the study aim to provide a new path to alleviate homelessness by unveiling the intersectional bias that affects PEH. Our research delivers a lexicon on homelessness, compiles an annotated dataset for homelessness and homelessness-racism intersectional (HRI) bias, evaluates LLMs as classifiers against these biases, and audits existing LLMs on HRI. Our goal is to contribute to homelessness alleviation by counteracting social stigma and restoring the human dignity of the persons affected.

During one night in 2024, 771,480 people in the United States were recorded as experiencing homelessness, the highest number ever documented [de Sousa and Henry, 2024]. These numbers continue to rise, not only in the United States but throughout the world. The last survey of homelessness from the United Nations found that 100 million people were homeless worldwide [Kothari, 2005]. Specifically, Nigeria, Pakistan, and Afghanistan each have over 4.5 million PEH [World Population Review, 2024]. This shows that the social challenge of homelessness persists across the world. The stigmatization of PEH negatively impacts the mitigation of homelessness in many ways, since the belief that the poor are undeserving reduces public support for mitigation policies [Applebaum, 2001]. Our research aims to contribute to the United Nations' first Sustainable Development Goal: Ending poverty in all its forms everywhere, by restoring the agency and capabilities of the PEH [Sen, 2001]. Recognizing the relationship between the stigmatization of PEH and efforts to reduce homelessness, we conduct research in NLP and LLMs to identify, classify, measure, and counteract stigma against PEH. Our work aims to answer the following research questions (RQs):

RQ1 - How does homelessness bias vary across US regions, and what factors influence this variation?

RQ2 - How well can existing LLMs classify stigmatization of PEH and HRI bias, and how can their accuracy and performance be improved to meet classification standards?

RQ3 - How biased are existing LLMs towards PEH (auditing)?

RQ4 - How can we counteract homelessness and HRI bias using LLMs and NLP?

2. AI Innovation: 2.1 Prior Work Our work builds on previous research by using NLP and LLMs to identify patterns in bias against PEH specifically, addresses the current limitations of LLMs in identifying and counteracting this type of bias. Existing related work presented an international comparative study on the criminalization of poverty in online public opinion [Curto et al., 2024], and a taxonomy of bias toward aporophobia [Rex et al., 2025]. These studies provided evidence of demeaning attitudes toward PEH [Ranjit et al., 2024]. However, previous studies on homelessness have been limited by lexicons containing a single word, 'homelessness,' and by collecting data from a single media source [Ranjit et al., 2024]. Moreover, previous studies do not establish clear correlations that explain the potential beliefs and biases against PEH.

2.2 Project Pipeline: To fill these gaps, we created an initial lexicon with the words: "homeless", "homelessness", "housing crisis", "affordable housing", "unhoused", "houseless", "housing insecurity", "beggar", and "indigent". We use the lexicon to scrape data from small and large counties in the United States that have similar socioeconomic demographic factors, including homelessness rates. To establish the potential association of homelessness bias with racism, we collect data from counties with different racial fractionalization scores to determine whether racial fractionalization contributes to PEH bias, as stated in the social science literature [Alesina and Glaeser, 2013]. After anonymizing the data, human domain experts annotate the data to clearly define what we are considering to be biased. The primary level classifies the text as "direct" or "reporting" bias [Rex et al., 2025] against PEH. The second level annotation includes stereotypes, known as OATH-Frames, such as 'not in my backyard', 'deserving/undeserving', and 'money aid allocation' [Ranjit et al., 2024]. This first annotated dataset serves as the "Gold Standard" against which LLM classifications will be compared to evaluate the ability of LLMs to detect homelessness bias. Using the results of the manual annotation and LLM classifications, we will develop metrics for HRI bias, using existing lexicons on racism. The results of the accuracy of the LLMs in detecting bias will indicate the feasibility of creating an HRI index. The index would allow the tracking of

homelessness and HRI bias over time, correlating it with social and political events, to inform homelessness-alleviation policy-making. In addition to evaluating how well existing LLMs can identify bias against PEH and HRI bias, we will also assess how well LLMs can mitigate it.

3. Transitional Evidence: This research has significant potential to address the socioeconomic impact of homelessness. The AI solution aims to flag and reframe negative language about homelessness, helping online communities become more inclusive. Additionally, the development of an HRI index will allow for the tracking of homelessness and related bias over time, providing data-driven evidence that can inform policy-making to alleviate homelessness. By revealing the intersectional bias affecting people experiencing homelessness. The project aims to provide data-driven evidence for nonprofits and local governments, strengthening campaigns to support people experiencing homelessness, and challenging dehumanizing narratives in everyday online discourse. Furthermore, it supplies actionable data on homelessness-related bias, informing municipal efforts and shaping broader anti-stigma policies. This comprehensive approach has the potential to foster more supportive online and offline environments for individuals experiencing homelessness. We have already used Llama

3.2 3B Instruct to classify Reddit data related to PEH. Llama

3.2 3B Instruct classifies the anonymized sentence 'I'm worried about a homeless shelter with addicts and people with mental issues in a residential neighborhood near a school' as 'not in my backyard' and a 'societal critique'. It also provides reasoning on its classification, and classifies it as not racist. We have collected data from cities across the United States and aim to classify and mitigate textual biases related to PEH from Reddit, X, news articles, and city council meeting minutes. One potential application of this is creating a Google Chrome extension that identifies and mitigates biases related to PEH. It would tell a user how text on a webpage is biased, offer a mitigated version of the text, and highlight the changes made.

4. RISE Dimensions: By focusing on identifying and mitigating bias against PEH, our research takes on the responsibility of addressing a critical social issue. It promotes inclusion by aiming to restore the human dignity of those affected by homelessness and counteracting social stigma. We recognize that the online data collected does not encompass the entire U.S. population, as 9.78 million people lack broadband access [Palmer, 2025]. Additionally, we only use a subset of online resources to analyze English textual data. This does not represent non-English speakers or other communication modalities. Ensuring the safety of handling data responsibly is critical. We anonymize the data to ensure that privacy is protected, and text cannot be traced back to individual users. Mitigating potential harms from biased language is also a key consideration. The project also incorporates ethical considerations. It acknowledges the risks associated with analyzing negative bias in online spaces when identifying and analyzing the negative bias in online spaces [Hovy and Spruit, 2016], especially with toxic language detection [Vidgen et al., 2019].

Track: Health AI and the Impact on Rural Healthcare

#16: Translational Research through a Developmental Lens: Integrating Dyadic and Longitudinal Methods to Inform Family-Based Interventions

Authors: Sarah Dennis, Lee Gettler and E. Mark Cummings

Problem Statement: An important focus within developmental science is the translation of research toward informing effective family-based interventions that are aligned with and supportive of children's development in the family context. This emphasis cuts across the fields of psychology, anthropology, and family studies, all integrated with a developmental lens. To understand how to promote the healthy development of children, especially those who are within vulnerable and underserved populations, it is vital that we also integrate perspectives from applied statistics to utilize appropriate methods that sufficiently capture developmental phenomena. This is necessary both when it comes to the basic research that informs applied implications, as well as the evaluation methods that are used to examine the efficacy of a given intervention. Yet, intervention efforts are often guided by and evaluated via cross-sectional studies or a focus on overall between-person differences. These approaches may not be appropriate for examining developmental processes, however. For instance, most work has focused on understanding child- and family-level traits associated with mental well-being in children. However, this is separate from understanding how children's mental health is affected when their relationship with their parents is more positive than usual; in order words, this question centers more around understanding the effects of parental states on children. The latter question would be more relevant to the development and evaluation of interventions that are designed to help gradually shift parenting behaviors or family dynamics in hopes of improving child outcomes. Furthermore, considerations such as timescale are relevant: Whereas many studies focus on yearly change, this does not provide implications on what can be done in the day-to-day to support child well-being. These are each different questions, but we often use methods that don't address these but still try to draw the same implications. Finally, it is important to consider how these processes play out within the family context by utilizing dyadic designs to accommodate parent-child and interparental dynamics.

Methodological Innovation and Translational Evidence:

Drawing from our own and other prior research, we underscore several existing longitudinal methods and showcase how these methods can be applied to the understanding of positive development in children. To this end, we present examples from several related papers (published and in progress). In the first set of examples, we highlight the importance of disaggregating between- and within-person effects (via multilevel modeling with person-mean-centering; Curran & Bauer, 2009; Wang & Maxwell, 2015) of parenting and family-level variables (Hoegler Dennis et al., in press). In the world of translational research, a goal is to be able to ascertain: "when a parent does X, a child's likelihood of positive outcomes increases, but when a parent does Y, a child's likelihood of negative outcomes increases. In other words, our goal is ultimately to understand the within-parent and within-family state-like processes that promote positive development and reduce maladjustment in children. Yet, often, results only reflect trait-level between-person processes: "Parents who do X tend to have children who have positive outcomes, and parents who do Y tend to have children who have negative outcomes. Given the number of examples of between-person effects that differ from within-person effects (Curran & Bauer, 2011), we present examples from prior work that disentangle these two types of effects and discuss the interpretations for family interventions.

We also present work that has examined these within- and between-person effects of parenting behaviors at daily and yearly timescales (Hoegler Dennis et al., under review(a)) through the application of dynamic structural equation modeling (DSEM; Hamaker et al., 2020). In many examples of developmental research, within-person effects are computed based on measures collected at the yearly-level. When it comes to parenting recommendations or programs, it is often the case that yearly-level findings are generalized and interpreted in terms of their implications for day-to-day parenting practices. However, to understand with greater specificity "when a parent does X on a given day, a child is more likely to experiences positive outcomes," it is necessary to disaggregate between- and within-level relationships utilizing daily-level data. DSEM (an approach that integrates components of multilevel modeling, structural equation modeling, and time-series modeling) provides the ability to disaggregate between- and within-person effects across multiple timescales.

Additionally, we illustrate ways of accommodating dyadic data to contextualize developmental processes in the family system by examining parent-child dyads (Hoegler Dennis et al., 2023; Vetterly, Hoegler Dennis, et al., 2024). For instance, by using dyadic growth curve modeling, we accommodate the nonindependence of dyadic data by: (1) specifying covariances between the latent intercepts and slopes across dyad members; (2) allowing dyad members' residuals at the same time to be correlated; and (3) permitting one dyad member's residual variances to be different from the other dyad member's residual variances (Planalp et al., 2017). By taking these steps, we can effectively model a dyads' interdependence of residuals, intercepts, and slopes and address research questions pertaining to the intercepts (e.g., initial, middle, or final levels, depending on how time is centered) of each dyad member, the changes (e.g., linear annual rates of change) for each dyad member, or covariances involving these latent variables.

Then, we demonstrate the integration and extension of these existing methods in novel ways. First, we integrate dyadic data into multilevel modeling with person-mean-centering and examine interactions between the between- and within-person effects of fathers and mothers on their children's adjustment (Hoegler Dennis et al., in press). We also combine between- and within disaggregation with time-varying interactions to examine the ways in which between- and within-effects of parents on their children can vary as children age (in prep(a)). Additionally, we show how DSEM can be expanded to incorporate dyadic data (Hoegler Dennis et al., under review (a)) and can also be extended to examine day-to-day reciprocal influences between parents (in prep(b)).

RISE Dimensions: Our goal throughout this presentation is to convey our responsibility for the ethical application of methods, interpretation of results, and translation of findings pertaining to correlates of positive development in children toward parenting recommendations, interventions, and even policy formation. To this end, we suggest that collaborative tutorial papers (Hoegler Dennis et al., under review (a and b)) may be avenues for encouraging responsible and ethical application of longitudinal methods toward family-based research. We also highlight how the implications of our findings can inform future directions for translational research, as well as the design and evaluation of interventions that support processes underlying children's positive development.

#33: AI-Driven Insights into Urban Architecture: Enhancing Health and Well-Being Across U.S. Cities with Responsibility, Inclusion, Safety, and Ethics

Authors: Siavash Ghorbany, Ming Hu, Siyuan Yao, Matthew Sisk, Chaoli Wang, Kai Zhang and Quynh Nguyen

The built environment profoundly shapes urban health, yet comprehensive, data-driven analyses across diverse U.S. cities have been lacking. This study addresses a grand societal challenge: understanding and mitigating the impact of urban architecture on mental, physical, and general well-being. We examine building characteristics across 19 major U.S. cities, spanning diverse climate zones and geographic regions, to identify influential factors and develop a predictive model applicable nationwide. Leveraging artificial intelligence (AI), we extracted architectural features—such as window-to-wall ratio and material types—from Google Street View imagery using Convolutional Neural Network (CNN) and large language models. These data were integrated with health metrics from the Centers for Disease Control and Prevention and analyzed using multiple regression and advanced machine learning models, including XGBoost, Support Vector Regression, Decision Trees, and Random Forests.

Our models achieved impressive R-squared values of 0.75, 0.76, and 0.82 for mental, physical, and general health, respectively, validated through testing in unseen cities (Seattle and Albuquerque). Key findings reveal that specific building features significantly influence urban health outcomes. Lead paint, affecting over 30% of urban households, emerges as a persistent hazard, strongly correlated with adverse health across all categories (coefficients: 0.063–0.097). Conversely, air conditioning prevalence is associated with improved health outcomes (coefficients: -6.316 to -7.641), highlighting the role of building quality. Traditional materials like wood and masonry, prevalent in older structures, correlate with better health compared to modern materials, while newer constructions show a negative association with well-being, possibly due to urban sprawl or material shifts.

Geographic variations underscore the need for context-specific urban planning. This research introduces an AI-driven innovation: a generalizable framework that harnesses large-scale visual data and machine learning to predict health outcomes based on architectural features. By integrating CNN-extracted building characteristics with health data, we provide a novel, scalable approach to urban health analysis, marking a significant advancement over prior single-city or survey-based studies. The methodology's robustness across climate zones and its testing in new urban contexts demonstrate its translational potential. Policymakers and urban planners can deploy this model to prioritize interventions—such as lead paint remediation or enhanced building standards—while architects gain insights for designing healthier cities. This research also integrates all these findings into a user-friendly web-based interface to demonstrate the United States health status in the urban areas at the census tract level for all types of users.

Aligned with the RISE framework, this study embodies Responsibility by using AI to inform equitable urban health solutions, Inclusion by focusing on diverse and underserved communities across the U.S., Safety by identifying and mitigating architectural health risks, and Ethics by addressing the unintended consequences of modern construction on well-being. The emphasis on historically underserved regions, where lead paint persists, amplifies diverse voices and ensures inclusive impact. By offering a predictive tool for smarter, healthier city planning, this work contributes to sustainable urban development and public health equity. Future research could expand to interior design factors or global contexts, but this study lays a foundation for responsible, inclusive, and ethical AI applications in tackling complex societal challenges.

#50: Early Warning Signals of the Ongoing Global Mpox Pandemic

Authors: Qinghua Zhao and Jason Rohr

Mpox, formerly known as Monkeypox, is a zoonotic viral disease primarily transmitted among humans through close contact. The case fatality rate ranges from 0.1% to 10%, and the virus is rapidly spreading to new regions worldwide. In 2024 alone, eight countries reported their first Mpox cases, having no prior recorded infections. To date, the virus has been detected in over 100 countries. While vaccination has the potential to mitigate its rapid spread, no specific Mpox vaccine currently exists. Smallpox vaccines are being used in practice; however, their availability remains limited. Therefore, it remains unknown if current vaccination efforts are enough to stop the global spread of Mpox. Moreover, it remains unclear whether Mpox outbreaks can be predicted during the pre-emergence phase, which regions are most vulnerable to future invasions, and what environmental and social factors drive its outbreaks.

To address these knowledge gaps, we applied Early Warning Signal and time-to-event analysis to daily case data across 129 countries from January 1, 2022, to February 15, 2025. I found that the mean lead time—the period between the detection of warning signals and outbreak onset—is 14.2 days. Regions with higher temperatures and a higher Human Development Index (HDI), an indicator of development and per capita income, exhibited longer lead times. In contrast, areas with higher human population density and greater precipitation experienced shorter lead times. Additionally, regions with lower HDI were more susceptible to Mpox invasion. The current use of Smallpox vaccines did not significantly extend lead times, and Mpox invasion success likelihood showed only a weak association with vaccination coverage. Our findings underscore the importance of integrating socio-environmental factors into global Mpox outbreak prediction, and indicate that current vaccination efforts alone are insufficient to curb the global spread of Mpox. Our study advances Early Warning Signal approaches for global pandemics, providing a foundation for a disease early warning system to guide proactive public health actions.

#73: A Responsible AI Framework for Hourly Seamless Ozone Estimation

Authors: Zhehao Liang, Stefano Castruccio and Paola Crippa

This study presents a high-resolution hourly surface ozone reconstruction framework developed using artificial intelligence (AI) techniques and multiple data sources. The framework incorporates AI-driven insight with interpretability and robust uncertainty estimation. The reconstruction methodology adheres to the principles of Responsibility, Inclusion, Safety, and Ethics (RISE). Cross-validation results demonstrate good spatiotemporal performance and robust accuracy in less densely populated land types, such as suburban areas.

#87: NGQA: A Nutritional Graph Question Answering Benchmark for Personalized Health-aware Nutritional Reasoning

Authors: Zheyuan Zhang, Yiyang Li, Nhi Le, Zehong Wang, Tianyi Ma, Vincent Galassi, Keerthiram Murugesan, Nuno Moniz, Werner Geyer, Nitesh Chawla, Chuxu Zhang and Fanny Ye

Diet plays a critical role in human health, yet tailoring dietary reasoning to individual health conditions remains a major challenge. Nutrition Question Answering (QA) has emerged as a popular method for addressing this problem. However, current research faces two critical limitations. On the one hand, the absence of datasets involving user-specific medical information severely limits personalization. This challenge is further compounded by the wide variability in individual health needs. On the other hand, while large language models (LLMs), a popular solution for this task, demonstrate strong reasoning abilities, they struggle with the domain-specific complexities of personalized healthy dietary reasoning, and existing benchmarks fail to capture these challenges. To address these gaps, we introduce the Nutritional Graph Question Answering (NGQA) benchmark, the first graph question answering dataset designed for personalized nutritional health reasoning. NGQA leverages data from the National Health and Nutrition Examination Survey (NHANES) and the Food and Nutrient Database for Dietary Studies (FNDDS) to evaluate whether a food is healthy for a specific user, supported by explanations of the key contributing nutrients. The benchmark incorporates three question complexity settings and evaluates reasoning across three downstream tasks. Extensive experiments with LLM backbones and baseline models demonstrate that the NGQA benchmark effectively challenges existing models. In sum, NGQA addresses a critical real-world problem while advancing GraphQA research with a novel domain-specific benchmark. Our codebase and dataset are available here.

#94: Health Technology Assessment framework of artificial intelligence in Radiology; Strategic decision makers` priorities.

Authors: John Olukuru, Miriam Miima and Joseph Onyango

Problem Statement: Enhanced diagnostic capacity and operational excellence in healthcare is advanced by development and adoption of artificial intelligence (AI) technologies in healthcare. Currently radiology in a pacesetter in AI technologies due to data abundance and pattern recognition.(1) Europe and North America have made bigger strides in the regulation and implementation of AI in radiology compared to Africa.(2,3,4) Assessment of software as a medical device locally remains unclear, which affects comprehensive standardization and articulation of its value proposition. Despite proven clinical effectiveness of AI in radiology, clinical translation remains opaque to end-users when technology is not articulated to the stakeholders.(5) Embedding user centered artificial intelligence in radiology promotes responsible, inclusive, safe and ethical AI deployment in healthcare. There is no explicit health technology assessment (HTA) framework in the systemic appraisal of economic, social, ethical and clinical health care priorities in Kenya. The study highlights perspectives from strategic decision makers on health technology assessment (HTA) priorities for AI in radiology.

AI Innovation: The strategic decision-making apex for AI implementation in radiology is made up of radiographers, radiologist, clinicians & health care managers, with governance support from policy makers & regulators, and knowledge management supported by academicians & researchers.(3,6) A health technology assessment framework derived from HTA core 3.0 model (7) with considerations from professional radiology multi-societies, was used to develop a local HTA tool.(2,3) Perspectives from 57 decision makers were analyzed through sequential elimination and multicriteria decision analysis. This action research supported iterations of the tool to a final interrater agreement of 0.98. The interclass correlation (ICC) using Cronbach alpha was acceptable for ethical domain (α=0.71) among other economic, social and clinical priorities.

Translational Evidence: Kenya like most African countries lags behind in the development of AI technology due to meager infrastructure and data poverty. This framework supports contextual decision making, with a holistic interrogation of social, ethical, economic and clinical priorities, during development and deployment of AI technologies in low resource set up. This framework primes policy recommendation in the regulation and technical requirements of local AI tools in radiology. RISE Dimensions: HTA framework promotes responsible development and deployment of AI systems through inquiry on clinical validation, patient safety, AI explainability and affordability. Multi stakeholder involvement fosters inclusivity in standardized collaboration, human centered design, transparent communication and education of AI and related concepts. Articulation of the value of AI using the HTA framework provides a reference for safe and ethical AI scored against technical, clinical, social and ethical inferences from the HTA core 3.0 model.

References

- Yardanova, M. Z. (2024). The Applications of Artificial Intelligence in Radiology: Opportunities and Challenges. European Journal of Medical and Health Sciences, 6(2), 11–14. https://doi.org/10.24018/ejmed.2024.6.2.2085 Boverhof
- BJ, Redekop WK, Bos D, Starmans MPA, Birch J, Rockall A, Visser JJ. Radiology AI Deployment and Assessment Rubric (RADAR) to bring value-based AI into radiological practice. Insights Imaging. 2024 Feb 5;15(1):34. doi: 10.1186/s13244-023-01599-z. PMID: 38315288; PMCID: PMC10844175.
- Brady AP, Allen B, Chong J, Kotter E, Kottler N, Mongan J, Oakden-Rayner L, Pinto Dos Santos D, Tang A, Wald C, Slavotinek J. Developing, purchasing, implementing and monitoring AI tools in radiology: Practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. J Med Imaging Radiat Oncol. 2024 Feb;68(1):7-26. doi: 10.1111/1754-9485.13612. Epub 2024 Jan 23. PMID: 38259140
- Kawooya MG, Kisembo HN, Remedios D, Malumba R, Del Rosario Perez M, Ige T, Hasford F, Brown JK, Lette MM, Mansouri B, Salama DH, Peer F, Nyabanda R. An Africa point of view on quality and safety in imaging. Insights Imaging. 2022 Mar 26;13(1):58. doi: 10.1186/s13244-022-01203-w. PMID: 35347470; PMCID: PMC8959275.
- Hua D, Petrina N, Young N, Cho JG, Poon SK. Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: A scoping review. Artif Intell Med. 2024 Jan; 147:102698. doi: 10.1016/j.artmed.2023.102698. Epub 2023 Nov 9. PMID: 38184343.
- Farič N, Hinder S, Williams R, Ramaesh R, Bernabeu MO, van Beek E, Cresswell K. Early Experiences of Integrating an Artificial Intelligence-Based Diagnostic Decision Support System into Radiology Settings: A Qualitative Study. Stud Health Technol Inform. 2023 Oct 20;309:240-241. doi: 10.3233/SHTI230787. PMID: 37869850
- Kristensen FB, Lampe K, Wild C, Cerbo M, Goettsch W, Becla L. The HTA Core Model®-10 Years of Developing an International Framework to Share Multidimensional Value Assessment. Value Health. 2017 Feb;20(2):244-250. doi: 10.1016/j.jval.2016.12.010. PMID: 28237203

#105: Empowering Caregivers Through AI: A Culturally Tailored Chatbot for Nutrition Support in Pediatric Cancer

Authors: Angélica García-Martínez, Beatriz Ribeiro Soares, Sisy Chen, Jane Stallman, Anna McCartan, Vik Miller, Fernando Sánchez, Horacio Márquez-González and Nitesh Chawla

Background: Caregivers of children with cancer experience high physical and emotional burdens, often worsened in low- and middle-income countries by poverty and inequities. These challenges compromise nutrition and caregiving, leading to poor treatment outcomes. To address this, the Lucy Family Institute for Data and Society developed SaludConectaMX, an mHealth system deployed at the Hospital Infantil de México Federico Gómez (HIMFG).
Methods: Participatory analysis with caregivers and health personnel identified barriers, misconceptions, and myths surrounding nutrition during chemotherapy. A nutrition chatbot was developed using St. Jude guidelines and fine-tuned in JSONL format for evidence-based, culturally tailored guidance. Validation included expert review, safety checks, and focus groups. Evaluation scenarios considered Mexican food practices, clinical safety, and caregiver usability.
Results: The chatbot demonstrated structural validity, cultural relevance, and clarity, with physicians confirming the accuracy of recommendations. Caregivers reported that the tool addressed their nutritional guidance needs and improved confidence in managing diets before, during, and after chemotherapy.
Conclusion: Fine-tuning the chatbot aligned it with HIMFG's educational and safety requirements. A Retrieval-Augmented Generation (RAG) system was implemented to further guide recommendations, utilizing a database of healthy, locally available foods to ensure safe, contextually relevant support for caregivers.

#120: Improving Health Through Housing Stability in the South Bend Community

Authors: Hannah Huston, Elizabeth Rhee, Melany Morales-Garibay, Siyuan Yao, Siavash Ghorbany, Matthew Sisk, Ming Hu and Chaoli Wang

Access to stable, affordable housing has become increasingly challenging in recent years, exacerbated by the COVID-19 pandemic and broader economic challenges. Research consistently shows that housing stability is a fundamental determinant of health. Yet, low-income households face growing risks of eviction, homelessness, and displacement. This study investigates the relationship between housing stability and health outcomes, specifically cardiovascular and mental health, in South Bend, IN, a city confronting a significant affordable housing gap. Our research evaluates how housing characteristics, such as physical conditions and affordability, influence health. We use computer vision models applied to Google Street View (GSV) imagery to analyze housing at scale to segment and classify property conditions. We leverage large language models (LLMs) to streamline data annotation, minimizing manual input and enabling scalable, cost-effective dataset generation. These annotated datasets train efficient vision models that identify indicators of instability, including inadequate insulation, structural disrepair, and signs of environmental neglect. We then integrate these assessments with health data using AI-driven analytics, modeling the relationships between housing conditions and health outcomes. Our models account for individual-level factors (age, gender, socioeconomic status) and community-level variables (poverty rates, racial composition, air pollution). This multifactor approach enables us to identify housing-related risks disproportionately affecting vulnerable populations, informing targeted and equitable interventions. Beyond analysis, the project explores passive mitigation strategies, such as retrofitting homes with shading devices and improved insulation, to enhance energy efficiency and indoor health without the high costs of new construction. Wearable health monitors and energy-use sensors will collect real-time data before and after retrofits, allowing us to measure intervention effectiveness and scalability. This research supports the RISE (Responsibility, Inclusion, Safety, Ethics) mission. AI tools are applied responsibly, emphasizing fairness, transparency, and community inclusion. Ethical practices underpin data collection, informed consent, and resident engagement. This work seeks to improve health outcomes and housing stability for underserved communities in South Bend and beyond by addressing systemic housing inequities and promoting affordable, data-driven interventions.

#141: A Bayesian Network Approach to Inferring Causal Relationships Based on Individual Behaviors in Aedes-Borne Diseases

Authors: Yuxin Meng and Alex Perkins

Problem Statement: Aedes-borne diseases, including dengue, chikungunya, Zika, and yellow fever, are illnesses transmitted to humans through the bites of mosquitoes from the Aedes genus. Aedes mosquitoes breed in containers and lay eggs in small collections of stagnant water. A key challenge arises in areas lacking consistent access to piped water and sanitation, where residents are often forced to store water, creating ideal breeding grounds for these vectors. At the present stage, we utilize household-level knowledge, attitude, and practice (KAP) survey data (published by García-Betancourt et al. 2015) collected in Girardot, Colombia, a city with a long history of water shortages and a widespread perception of expensive charges for water. The KAP survey captures each individual's socioeconomic characteristics (e.g., age, gender, monthly family income), their knowledge of various dengue prevention measures (e.g., knowledge of dengue transmission, knowledge of washing pools and tanks to prevent mosquito breeding, knowledge of how to use bed nets), and their actual intervention or preventive behaviors (e.g., whether they do indoor fumigation, whether they cover water containers, whether they add chemicals in water). We investigate the determinants of these behaviors, identify how diverse cues may lead to varied behavioral responses, and determine what socioeconomic characteristics cause the differences in prevention knowledge by applying Bayesian network learning. This machine learning approach reveals the reasons for the occurrence of individual behaviors and the existence of individual knowledge regarding Aedes-borne diseases.

AI Innovation: Bayesian network learning represents a novel application in KAP survey analysis. Unlike simple correlational statistical analyses, Bayesian networks capture conditional dependencies among socioeconomic factors, individuals' prevention knowledge and preventive behaviors and enable us to infer not only the causal structure among variables but also the conditional probability distributions of each variable. Moreover, despite the presence of skipped questions, unanswered items, or subjective attitude scales in the KAP survey data, Bayesian network learning can handle such imperfections. In other words, this approach naturally accommodates both uncertainty and incomplete data, enabling more reliable and robust inferences even under imperfect survey conditions. Strong causality consistently emerged after extensive resampling, leading to the construction of a stable causal network. We applied bootstrap resampling and learned 1,000 network structures using the hill-climbing algorithm. A predefined blacklist was incorporated during structure learning to enforce plausible causal assumptions. Specifically, we prohibited causal flows from intervention behaviors to prevention knowledge, as well as from knowledge or behaviors to socioeconomic factors, ensuring that the directionality of edges aligns with logical expectations. We then constructed a consensus network by retaining only those edges that appeared in a high proportion of the sampled networks and maintained consistent directionality. This process yielded a stable and interpretable structure for subsequent modeling.

Translational Evidence: Our current results support a causal pathway that indicates that socioeconomic factors determine individuals' knowledge of prevention, while it is the knowledge that causes various prevention behaviors. The result is shown in Fig 1. This figure highlights that, even if we only know individuals' socioeconomic status, we cannot infer their prevention behaviors unless we have more information regarding their prevention knowledge. Specifically, monthly family income appears to be a determinant of an individual's knowledge regarding the avoidance of stagnant water and the prevention of mosquitoes breeding in it. As a result, both the removal of unnecessary containers and the practice of storing water depend on people's awareness of preventing stagnant water. This hierarchical network will further assist us to understand human behaviors that affect individual risk of Aedes-borne diseases. Fig 1: Causal pathway by Bayesian Networks. In our subsequent research, the conditional probabilities obtained through Bayesian network learning will contribute to defining the decision-making rules within an agent-based model. Each household agent will be initialized with attributes, such as socioeconomic characteristics, levels of prevention knowledge, and prevention behaviors. Ultimately, this approach will bridge micro-level behavioral modeling and macro-level epidemic outcomes, providing actionable insights for designing targeted and equitable public health interventions to control Aedes-borne diseases.

RISE Dimensions: Aligned with Inclusion AI, our modeling significantly promotes inclusiveness by incorporating heterogeneous individual-level data from underserved communities with limited water and infrastructure rather than relying solely on national or city-wide macro-level data. Individuals traditionally marginalized by infrastructural limitations are accounted for in our research through the data used to inform it. Our results will furthermore contribute to designing agent-based simulations that will aim to inform equitable and culturally sensitive public health interventions, minimizing potential harms and maximizing benefits for vulnerable populations. Through this commitment to Inclusive AI, our work demonstrates how AI (machine learning) techniques can be harnessed to empower marginalized communities and support the development of fairer, more effective disease control strategies. García-Betancourt, Tatiana, et al. "Understanding water storage practices of urban residents of an endemic dengue area in Colombia: perceptions, rationale and socio-demographic characteristics." PloS one 10.6 (2015): e0129054.

#142: An Artificial Intelligence Tool for Patients and Clinicians to Improve Utilization of Medication Access Programs

Authors: Yejin Seo, Noha Keshk, Alexandros Psomas, Faria Chaudhry, Jasmine Gonzalvo, Rakhi Karwa and Alan Zillich

Problem Statement: Medication insecurity refers to the inability to pay for prescribed medication, a growing issue affecting 40% of U.S. adults taking four or more prescriptions. This insecurity is driven by high medication costs, unemployment, competing financial priorities, and limited insurance coverage. Rising medication costs contribute to poor medication adherence in 50-60% of patients, leading to worsened chronic disease outcomes, 125,000 annual deaths, and significant healthcare expenditures. The concept of medication insecurity highlights the need for equitable access to necessary medications for all individuals. While 89.5% of patients express interest in cost-reducing tools, many healthcare providers lack the knowledge and resources to address affordability concerns, with only 12% utilizing discounted drug lists and other potentially available resources. Ubiquitous AI sources, such as ChatGPT, often provide inapplicable suggestions for discounted medication resources, highlighting the need for the development of a new Artificial Intelligence (AI) model that relays more accurate and relevant recommendations.

AI Innovation: A proof-of-concept Large Language Model (LLM) model using Application Programming Interface (API) tools from ChatGPT was developed to explore the feasibility of affordable medications resource recommendations based on user-specific inputs such as drug name, dose, zip code and insurance type. As a first step, 50 patient cases were collected to test the accuracy of the pilot LLM. Each case contains medication name, dose and quantity, ZIP code, and insurance type as well as medication access recommendations related to each case, providing contextual information essential for modeling real-world affordability scenarios. These cases are structured with individual text rows and categorical labels and can serve as "inputs" and "outputs" for LLM training and tuning. Manual validation was conducted on a subset of records to assess completeness and consistency, and feedback from experts from the Center for Health Excellence, Quality and Innovation (CHEQI) was used to iteratively refine both the dataset and model outputs. Our goal is to further fine-tune and prompt engineer the LLM to make better recommendations. A major challenge is testing the accuracy of the LLM. Patient cases will be repeatedly obtained and the LLM refined until near-perfect accuracy of recommendations can be guaranteed, after which point a working prototype LLM will be soft-launched. The prototype model will be implemented using publicly available API based LLM frameworks and accessed through a lightweight interface designed for ease of use by healthcare workers or patients. This tool will be validated by healthcare providers to ensure the accuracy of the recommendations from the LLM model prior to being made available for patient use. Then, this tool will be pilot tested at partnering health systems that serve patient populations in both urban and rural communities of Indiana who face challenges with medication affordability. Future iterations of the system will incorporate user feedback to improve usability and boost utilization of the model.

Translational Evidence: Our objective is to establish a fully functional, user-facing medication access (MART) LLM-powered tool within 24–36 months. The pilot phase of this project is expected to demonstrate the feasibility of using a large language model to generate personalized medication resource recommendations based on user-specific inputs, such as drug name, dose, zip code and insurance type. The first 6 months will focus on optimizing data pipelines, expanding the dataset, and refining the model's interface and outputs. These early stages will also help identify key areas for improvement in the model's interpretability, factual accuracy, and data structuring. By months 12–18, a minimally

viable product (MVP) will be deployed and evaluated through pilot testing in collaboration with health systems serving urban and rural Indiana. This early deployment will allow us to test the model's effectiveness in real-world settings and gather critical user feedback. By the 24-36 months, controlled testing and validation of MART-LLM is planned for completion. Its recommendation will be compared against expert generated and real-world patient navigation outcomes. We aim to reach and support over 1000 patients across three health-systems, with special focus on socioeconomically disadvantaged and medication-insecure populations in Indiana. Our long-term goal is to boost user engagement tenfold by integrating cost-reduction programs and tailoring the tool to improve perceived usability. Utilization metrics will provide essential feedback for ongoing model refinement. Incorporating patient and provider feedback will ensure continuous improvements in the tool's accessibility, equity, and clinical relevance.

RISE Dimensions (Responsibility, Inclusion, Safety, Ethics) The development of MART-LLM aims to improve medication access for all patients, with a focus on benefiting socioeconomically disadvantaged populations. By providing resources for more affordable treatments, MART-LLM supports improved medication adherence, leading to better health outcomes, and reduced overall healthcare costs. This tool aims to address current gaps in medication affordability, contributing to a more equitable and efficient healthcare system. From a safety perspective, the model does not collect any personally identifiable patient information, protecting user safety and privacy. All data utilized to train the model is either deidentified real-world patient cases or mock-patient cases. Our plan to introduce the LLM model to healthcare workers first, prior to its availability for patient use, will help ensure adequate validation of recommendations and would safeguard patient safety by keeping the "clinician in the loop". Additionally, if the tool generates a recommendation of medication changes, the patient will be referred to their healthcare care providers to discuss any potential changes of the medication, safeguarding against inappropriate or harmful intervention.

#148: Prediction System for Freezing of Gait in People with Parkinson's Disease

Authors: Lourdes Martinez-Villaseñor, Ari Barrera-Animas and Hiram Ponce

Parkinson's disease (PD) is a progressive neurodegenerative disorder that significantly affects motor control, leading to symptoms that drastically reduce patients' quality of life. One of the most disabling motor symptoms is freezing of gait (FOG), which is characterized by brief, episodic inability to start or continue walking despite the intention to move. FOG is common in advanced stages of PD and is closely associated with falls, loss of independence, and psychological distress. Current pharmacological and physical therapy treatments for FOG are often ineffective or lose efficacy when not applied at the right moment, underscoring the urgent need for more reliable, real-time interventions. This proposal aims to develop a predictive system capable of identifying FOG episodes a few seconds before they occur. The system will use wearable sensors and camera-based video capture to collect motion data, which will be processed using real-time machine learning algorithms. By detecting precursors of FOG, the system can deliver auditory stimuli—timed specifically to each patient's needs—to help them regain normal gait and avoid freezing events. The research integrates multidisciplinary expertise, combining artificial intelligence, neuroscience, and rehabilitation science to create a clinically relevant solution.

The methodology consists of five key stages: (1) design of the experimental protocol, including activity planning, participant recruitment, and data acquisition setup; (2) ethical approval by research committees; (3) data collection during controlled sessions with Parkinson's patients under medical supervision; (4) development, training, and validation of machine learning models for FOG prediction; and (5) real-time testing of the predictive system to evaluate precision, sensitivity, and response time. Collaborating institutions include Universidad Panamericana, Clínica Coyoacán, and Actipulse Neuroscience, with support from experts in AI, neurology, psychology, and sensor-based technologies. Currently, most research focuses on detecting FOG rather than predicting it. Only two prior studies have explored FOG prediction—one using statistical methods and another based on inertial sensor thresholds—both with preliminary results. To date, no known research has applied machine learning techniques using wearable and vision-based technology to build predictive models for FOG.

This project therefore represents a novel and significant contribution to both the scientific and medical communities. An ethically grounded design is essential in the development of health-related technologies, particularly those involving vulnerable populations such as individuals with Parkinson's disease. In this project, ethical considerations have been

integral from the outset, with a focus on ensuring privacy, non-maleficence, accountability, transparency, fairness, and explainability. Privacy is protected through secure handling of sensitive data, including video recordings and physiological signals, which are stored locally and never uploaded to the cloud. The principle of non-maleficence guides all experimental protocols to minimize physical and psychological risks to participants. Accountability is upheld through oversight by institutional ethics committees, informed consent procedures, and well-defined roles among interdisciplinary team members.

Transparency is ensured by openly communicating the project's objectives, methods, and findings with participants and stakeholders. Fairness is addressed through inclusive experimental design that accounts for individual differences and avoids introducing bias in predictive models. Crucially, explainability is emphasized to ensure that the machine learning models used to predict freezing of gait produce interpretable results, enabling clinicians and patients to understand and trust the system's recommendations. Together, these ethical principles foster responsible, human-centered innovation in digital health technologies. In conclusion, this research addresses an urgent clinical need with a technologically innovative and patient-centered approach. By combining wearable and visual sensing with intelligent prediction algorithms, the project seeks to deepen our understanding of Parkinson's disease while providing a practical tool to improve the daily lives of those affected. The outcomes of this work could pave the way for future personalized therapies and smart assistive technologies for Parkinson's patients worldwide.

#158: A Taxonomy-Guided Evaluation of Sociolinguistic Diversity in Swahili NLP

Authors: Kezia Oketch, John Lalor and Ahmed Abbasi

Problem Statement: Despite the success of large language models (LLMs) in advancing natural language processing (NLP), their benefits remain unequally distributed across languages and cultures. Most NLP systems are trained and evaluated on standardized, high-resource language varieties, overlooking the rich sociolinguistic diversity present in real-world language use. This gap disproportionately affects low-resource languages like Swahili, where linguistic variation, including code-mixing, Sheng (urban youth vernacular), tribal lexical influence, and loanword integration, is often treated as noise and removed during preprocessing. This leads to systematic model failures that misrepresent underrepresented groups, especially in high-stakes domains such as healthcare communication and public services. The inability to model linguistic diversity risks further marginalizing these communities in algorithmic systems and impeding equitable AI adoption. This grand challenge highlights the urgent need for AI systems that recognize and equitably model linguistic variation rather than marginalizing it.

AI Innovation: We develop a new culturally grounded Swahili NLP testbed focused on four health-related psychometric tasks: health literacy, trust in doctors, anxiety visiting the doctor, and health numeracy. Our innovations include:

Dataset Creation: We adapted and translated validated English-language survey items from prior work measuring four health-related dimensions: Anxiety Visiting a Doctor, Subjective Health Literacy, Health Numeracy, and Trust in Doctor. Translation into Swahili was performed by a certified translator, and then independently reviewed by two Swahili language experts for cultural fluency and contextual accuracy, with discrepancies resolved through consensus. To ensure broad demographic representation, we partnered with GeoPoll, a mobile-first crowdsourcing platform with deep reach in Kenya. GeoPoll enabled us to access rural and urban populations across tribal, educational, and generational lines. We conducted a 125-participant pilot to validate the survey design and assess feasibility. The final dataset comprises 2,170 free-text Swahili responses (one per psychometric task per participant), paired with survey-based psychometric scale responses and detailed metadata on age, gender, income, education, and tribe. We implemented best practices for data quality assurance, including attention checks, manual cleaning, and incentive-based compensation, ensuring high-quality sociolinguistically rich data.

Taxonomy Development: Based on a grounded linguistic analysis of the dataset, we developed a structured taxonomy of sociolinguistic variation in Swahili. This taxonomy was derived using a mixed-methods approach: frequency-based extraction of candidate lexical items, native speaker annotation, and iterative refinement with expert linguistic input. The subset categories included in the taxonomy are: code-mixing (Swahili-English blends), Sheng (urban youth vernacular),

tribal lexical, and loanwords. Each text was manually annotated for the presence of these features, generating count indicators used in downstream modeling.

Fairness-Aware Evaluation Framework: We evaluated four multilingual PLMs (SwahBERT, XLM-RoBERTa, AfriBERTa, and mBERT) and four instruction-tuned LLMs (Llama 3.1-405B, Llama 3.1-8B, Qwen2.5-72B, and Qwen2.5-7B) on their ability to predict ground-truth scores for each psychometric task. We computed response-level prediction errors (Δ = actual − predicted) and modeled these errors as functions of both linguistic features and demographic attributes. We used OLS regression, and computed group fairness metrics including Disparate Impact (DI), Fairness Violations (FV), and ΔxAUC, enabling a principled, interpretable fairness audit at scale. This methodology enables an interpretable, systematic evaluation of how language variation impacts model behavior, moving beyond traditional accuracy-only assessments.

Translational Evidence: Our findings show that LLMs and PLMs consistently underperform on texts with high linguistic variation, with model errors disproportionately affecting speakers using Sheng, tribal lexicon, or code-mixed language. These results have multiple pathways for real-world translation: Improved AI robustness in multilingual systems: By modeling linguistic variation as signal, our framework enables LLMs to generalize more reliably across sociolinguistically diverse populations. This has immediate implications for improving NLP robustness in multilingual deployments, where speaker variation is the norm, not the exception. Integrating sociolinguistic features as auxiliary inputs or training signals can improve model calibration and reduce overfitting to standardized forms of language. Fairer deployment in health, education, and public services: Public-sector NLP applications such as patient triage bots, literacy assessments, or chatbot-based health counseling, often rely on text inputs from diverse speakers. Our results show that current models systematically misjudge the quality or intent of responses written in non-standard Swahili. Adapting NLP tools using our annotated corpus and fairness metrics enables more inclusive design and reduces the risk of exclusion or miscommunication in critical domains. For instance, healthcare systems serving Swahili-speaking populations could use our data and methods to train more representative triage classifiers.

Policy and governance implications for AI equity: Our work provides a template for regulators, standards organizations, and AI practitioners seeking to evaluate equity in language technologies. By demonstrating how demographic and linguistic dimensions can be systematically incorporated into fairness audits, our framework contributes to the growing movement for AI impact assessments. In contexts such as Kenya's AI policy development or WHO health literacy initiatives, our annotated dataset and methodology offer a deployable, data-driven approach for monitoring equitable AI outcomes. Building models that honor sociolinguistic realities strengthens socio-economic participation and representation for African communities and beyond. **RISE Dimensions:** This research directly supports the goals of Responsible, Inclusive, Safe, and Ethical (RISE) AI development:

Responsibility: We treat linguistic diversity not as a nuisance variable but as an essential property of real-world language use. Our error modeling framework identifies systematic model failures tied to sociolinguistic traits and provides transparent evaluation pathways grounded in equity.

Inclusion: Our dataset prioritizes representation across tribal, regional, and socioeconomic groups, and our taxonomy centers the lived linguistic realities of Swahili speakers. We demonstrate that ignoring these factors leads to performance disparities, underscoring the need to design with inclusion from the ground up.

Safety: Disaggregated model auditing reveals fairness risks that would be invisible under traditional aggregate metrics. For example, we identify model overconfidence and misprediction concentrated among texts written by speakers using Sheng, or those from specific tribal groups. Our findings offer early-warning signals for downstream harms in sensitive contexts such as health information triage or social service eligibility screening.

Ethics: By surfacing model behavior linked to culturally significant language use, we move away from language universalism and toward contextual AI that respects identity. Our framework challenges the epistemic bias that standardization equals quality, proposing instead, a paradigm where linguistic difference is valued and protected in AI systems. This project contributes to the broader vision of AI technologies that are not only high-performing but also accountable, culturally aware, and just, especially for populations historically left out of NLP research and development.

#165: A Novel AI-powered Pipeline for Alzheimer's Disease Classification using Spontaneous Speech and Vector Embeddings

Authors: Daniel Zhou

Problem Statement: Alzheimer's disease (AD) represents a growing global public health challenge. Early diagnosis of Alzheimer's is crucial for a timely intervention and effective care planning. Traditional diagnosis methods, such as cognitive and behavior tests, brain imaging, biomarkers & lab tests, often come with significant limitations, including reliance on medical facilities, invasive and expensive tests, and the risk of missing symptoms in the early stage. It is essential to explore alternative methods that offer a more accessible and non-invasive approach to AD diagnosis and screening.

AI Innovation: Spontaneous speech, naturally occurring in real-world settings, holds potential as a digital biomarker for accessible and non-invasive AD diagnosis and screening. While acoustic and linguistic features have been extensively explored, the existing studies often rely on domain knowledge and manually crafted transformation.

Large Language Models, trained on vast datasets, can interpret the semantic meanings behind words, sentences or images. Text can be transformed into an embedding - a vector (list) of floating point numbers in high dimensional space. The distance between two vectors indicates their relatedness, with small distances signifying strong relatedness and large distances suggest weaker ones. There is a huge potential in using AI and especially the large language model for uncovering the latent semantic meanings within the speech data.

In this study, we developed an AI-powered pipeline to automatically process and transcribe spontaneous speech recordings, using AssemblyAI API (with utterance filter support) and OpenAI Whisper API (added post process with LLM prompts to filter out speakers). The dataset used in this research is the spontaneous speech recordings from the Alzheimer's Dementia Recognition through Spontaneous Speech (audio only) - the ADReSSo Challenge 2020. The dataset features speech recordings and transcripts of spoken picture descriptions based on the Cookie Theft picture from the Boston Diagnostic Aphasia Exam, including 78 participants with Alzheimer's disease (AD) and 78 non-AD participants.

As a baseline, we extracted 41 acoustic features using the Python Librosa library; Additionally, we applied OpenAI Vector Embedding models, both text-embedding-3-small with 1536 dimensions and text-embedding-3-large with 3072 dimensions, to extract semantic features from transcription file. Using acoustic features from audio files and semantic features from vector embeddings, we trained classification models to distinguish AD patients from healthy individuals, as well as regression models for predicting Mini-Mental State Examination(MMSE) score.

For AD classification, we implemented 6 classification models using scikit-learn library: Logistic Regression(LR), Random Forest Classifier (RF), Support Vector Classifier (SVC), XGBoost Classifier (XGBoost), K-Nearest Neighbors Classifier (kNN) and Muti-Layer Perceptron classifier (MLP). With acoustic features only, the best model precision=0.748 and best accuracy=0.66. With the vector embedding approach, best results can be achieved with the provided manual transcription (filtered for the "PAR" user) for precision=0.911 and accuracy=0.883. For the automated transcription results, we obtained better results (precision=0.874 and accuracy=0.829) with AssemblyAI transcription with speaker diarization support, compared to those with Whisper transcription (precision=0.774 and accuracy=0.725). Furthermore, we improved the Whisper model performance by adding post-processing steps using GPT-4o LLM prompt engineering (precision=0.868 and accuracy=0.819). Among the 6 models, the Support Vector Classifier (SVC) consistently outperformed other classification models.

For MMSE score prediction, we trained Ridge Regression, Random Forest Regression, and XGBoost Regression models. The RandomForestRegressor based on AssemblyAI transcript outperformed other regression models, with the Root Mean Squared Error (RMSE) = 4.51 and R-squared (R2) = 0.45. In summary, we have successfully developed an AI-based pipeline to automate spontaneous speech analysis, and demonstrated that LLM based vector embedding is a viable approach for Alzheimer's classification and MMSE score prediction.

This AI-based pipeline aligns with the RISE dimensions:

- Responsibility: this AI pipeline has the potential to assist healthcare professionals in AD early detection, which can lead to timely intervention and better patient outcomes. We will also evaluate its effectiveness for other speech-sensitive conditions such as aphasia, depression, PTSD, and Parkinson's disease.
- Inclusion: The LLM vector embedding model can be applied to non-English spontaneous speech recordings, which are also available in DementiaBank. Broadening the language support ensures access for diverse linguistic communities and enhances healthcare equity for underserved populations.
- Safety: By automating audio transcription and leveraging AI for detection, the pipeline minimizes human error and inconsistency in manual assessments. This application can provide a scalable, accessible, and standardized AI-driven approach, reducing misdiagnoses and ensuring that patients receive accurate evaluations.
- Ethics: Ethical considerations are central to AI deployment in healthcare. The data contribution and retrieval of DementiaBank follow a strict IRB approval process, with informed consent from individual participants for data-sharing. The speech recordings collected from diverse age and gender groups helps to avoid model discrimination.

Track: Human-Centered Responsible AI

#28: From Utilitarianism to Pragmatism: Evaluating AI-Generated Ethical Reasoning through Diverse Modeling Techniques

Authors: Laura Duparc and Alexander Belikov

This project addresses the grand challenge of creating trustworthy, context-aware AI systems capable of replicating nuanced ethical reasoning drawn from classical philosophical traditions. In response, we propose an innovative methodology for generating and evaluating philosopher-inspired personas within Large Language Models (LLMs). Our approach constructs two distinct personalities—one grounded in John Stuart Mill's Utilitarianism and the other in William James's Pragmatism—by systematically comparing two training strategies (knowledge graph–based modeling versus conventional fine-tuning) alongside two model architectures (reasoning-enhanced versus regular models). This approach produces four distinct configurations: one based on knowledge graph–based modeling with a standard architecture; one based on knowledge graph–based modeling augmented with reasoning capabilities; one derived from conventional fine-tuning using a standard architecture; and one derived from conventional fine-tuning enhanced with reasoning capabilities. By presenting both simulated philosopher-agents with an identical set of ethically charged questions, we systematically compare their responses using expert evaluations, as well as quantitative similarity measures that assess overlap and divergence between models trained with and without a knowledge graph. Additionally, each model critiques its counterpart's fidelity to the original philosophical doctrines, while an independent AI assistant (ChatGPT) provides a parallel external evaluation. Our methodology not only advances AI innovation in replicating deep intellectual traditions but also offers translational evidence with significant socio-economic and policy implications. By rigorously benchmarking diverse LLM architectures, our project informs strategies for deploying AI systems in educational settings, public policy decision-making, and digital platforms—areas where responsible, inclusive, safe, and ethical reasoning is paramount. This use of interdisciplinary framework aims to allow to amplify diverse intellectual voices and promotes transparent AI practices.

#55: All-Female Teams Drive More Disruptive Ideas in Science

Authors: Nandini Banerjee and Diego Gomez-Zara

Disruptive research challenges conventional wisdom, reshapes disciplines, and paves the way for new scientific directions. A classic disruptive scientific work is Watson and Crick's discovery of the DNA double helix, which revolutionized biology and numerous other fields and also had an immense social impact. Recent advancements in the science of science provide a novel metric to quantify the transformative nature of scientific work called the disruption score1,2. Understanding disruption can help policymakers, institutions, and researchers improve how we evaluate scientific research, allocate resources, foster collaboration, and accelerate breakthroughs that drive societal progress.

Previous research has analyzed the characteristics of scientific teams that produce disruptive papers. For example, smaller scientific teams1, newer collaborative relationships among team members3, egalitarian team structure4, and teams working in proximity5 have been shown to be disruptive. Despite these findings, the influence of a team's gender composition on disruption remains uninvestigated. Although there is still an underrepresentation of women in academia (e.g., graduate degrees, active authors, tenure positions), their participation in research has been rising, leading to a shift in team compositions from all-male to gender-diverse research teams.

To examine how these shifting team dynamics have affected the quality of scientific output, we analyzed over 130 million papers published between 1950-2019, their authors, their publication texts, and their disruption scores obtained from the SciSciNet dataset6 and Microsoft Academic Graph. The disruption (D) score varies from -1 to 1, where a value of -1 denotes papers that consolidate their predecessors' ideas, and a score of 1 denotes papers that challenge their predecessors' ideas. We find that all-female teams are at the forefront of driving highly disruptive breakthrough ideas in science (Fig 1a). All female teams have produced the most significant proportion of disruptive (the top 5% of D score) work annually compared to all-male (two-sided Wilcoxon signed-rank test, $W=2{,}377$, $p<0.001$, Cliff's Delta = 0.5) and

balanced, i.e., 50-50 teams (W=2,485, p<0.001, Cliff's Delta = 0.9). We find that this observation holds across different time periods, team sizes, citation counts, author career lengths, and fields. To investigate this gender composition-based difference in disruption, we examined the titles and abstracts of highly disruptive papers published by both all-female and all-male teams. We used the pre-trained language model SPECTER27 to obtain text embeddings. This model utilizes both paper titles and abstracts to generate the embeddings. We used UMAP8 to plot the embeddings in a two-dimensional space.

We selected three representative fields in science and engineering: medicine, computer science, and chemistry. We find that in certain fields, all-female teams have explored topics that have not been extensively examined by all-male teams. In medicine, for example, we find that all-female teams have investigated underrepresented gender-related research topics centered on female health and gender (Fig 1b). We find gender-based distinctions in research topics in computer science, where all-female teams show noticeable differences in research interests (Fig 1c). Moreover, in highly specialized fields such as chemistry, gender-driven disruption may occur independently of the research topic (Fig. 1d). Overall, scientific works addressing research gaps are naturally more prone to be disruptive as they bridge missing links in science and, thus, create new research directions in their fields. Thus, our study highlights the critical role played by all-female teams in the exploration and discovery of novel research directions, emphasizing the importance of recognizing and valuing their contributions to modern science. Our findings also provide empirical evidence for institutions and policymakers to design more inclusive evaluation and funding mechanisms. By fostering inclusivity and facilitating the exchange of diverse ideas, all-female teams can sustain their role in pioneering innovations and advancing scientific discovery.

#56: Innovation and Conservation: Assessing AI's Water Footprint and Its Role in Sustainable Resource Management

Authors: Annie Chen and Alexi Orchard

The environmental impacts of Artificial Intelligence (AI) development and usage, including its damaging effects on the climate crisis, can no longer be ignored. As the literature in this space continues to develop, this paper aims to support future research by examining AI's environmental trade-offs. An analysis of such trade-offs would help uncover an ideal path – prioritizing conservation, fostering innovation, or finding a balanced integration of both – to a sustainable future. While AI has revolutionized many sectors and industries, its expansion increases environmental strain, particularly on water resources. "Even considering the lower estimate, the…water withdrawal of global AI is projected to reach 4.2-6.6 billion cubic meters in 2027, which is more than the total annual water withdrawal of Denmark or half of the United Kingdom" (Li 2023). Despite these challenges, AI promises to improve urban systems, waste management, and energy efficiency. These innovations illustrate possible opportunities to develop climate mitigation solutions. AI's duality as both a potential benefit and harm has been subject to ongoing research, such as water demand forecasting, desalination, and resource optimization. This survey paper synthesizes recent research on the impacts of generative AI on water consumption specifically. In the context of the United Nations Sustainable Development Goals (SDGs), such as SDG 6 (Clean Water and Sanitation), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 13 (Climate Action), we explore how AI-powered technology – such as predictive modeling, smart irrigation systems, and waste management frameworks – can enhance conservation efforts. We also assess AI's water and energy footprint, highlighting the need for ethical and sustainable AI applications and infrastructure. Academics, experts, and policy makers across domains – including public health, developmental economics, structural sociology, and climate law – remain divided on how to address this crisis. Some argue that society cannot continue relying on technologies to save us when such interventions exacerbate the water crisis (Shiva 2002), while others believe that social systems and innovation can transform ecosystems, improving our Earth's carrying capacity through human ingenuity (Ellis 2013). This debate has only intensified, as "by 2050, an additional one billion people are expected to live with extremely high water stress," even in the most optimistic scenario (Kuzma 2023). With such threats to human and ecological prosperity, adopting AI is only advisable if the ecological benefits outweigh the costs. AI, as a sustainable solution or an environmental strain, requires further interdisciplinary inquiry. We aim to support continued research by consolidating literature on AI applications in water management, examining its water footprint alongside "green AI" – environmentally sustainable and energy efficient technologies

(Verdecchia, et al. 2023). By considering both technological development and environmental constraints, we aim to foster a balanced discourse on AI's role in fostering climate resilience. We anticipate that this research will inform stakeholders and policymakers in navigating the relationship between innovation and conservation.

- Ellis, Erle. "Overpopulation Is Not the Problem." The NewYork Times, 2013.
- Kuzma, S. (n.d.). 25 countries, housing One-Quarter of the population, face extremely high water stress. World Resources Institute. Link.
- Li, Pengfei, et al. "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models." arXiv (Cornell University), Jan. 2023, doi:10.48550/arxiv.2304.03271.
- Shiva, Vandana. Water Wars: Privatization, Pollution, and Profit. Cambridge: South End Press, 2002.Verdecchia, Roberto, et al. "A Systematic Review of Green AI." Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, vol. 13, no. 4, June 2023, doi:10.1002/widm.1507.

#75: Unleashing the Metaverse's Potential for Collaboration and Innovation

Authors: Diego Gomez-Zara, Mariana Fernandez, Yunhao Xing and Niloofar Sayadi

Problem Statement: The metaverse promises to revolutionize the ways we work, learn, and interact with others [1, 2]. This technology—which is already reshaping games and industry—allows people to meet in immersive virtual spaces, adopt different identities, record actions, and trade digital objects. Augmented, Virtual, and Mixed Reality (AR/VR/MR) technologies are enhancing remote social interactions and providing ever more realistic immersive experiences [3, 4]. As a result of the extraordinary recent advances in hardware, software, and networking, the metaverse has become a viable option to facilitate remote work.

Despite the metaverse's immense potential, organizations have struggled to understand and use this technology. The lack of metaverse applications and appropriate designs for groups has been the biggest barrier to metaverse adoption. The research community needs better knowledge of the ideal design of technological configurations and affordances to facilitate team science's use of the metaverse in the future. To test these opportunities and challenges, we will examine the mechanisms that may allow the metaverse to become a productive tool for collaborations, as well as its effects. Our initial research questions are:

RQ1: What are the main advantages and disadvantages of using the metaverse for tasks compared to meeting in person or on video-conferencing platforms?

RQ2: Can the metaverse enable human users to collaborate with AI agents embodied as avatars?

The goal of this project is to design, build, and evaluate the impact of metaverse technologies on scientific work. First, we will run a laboratory experiment to evaluate the effectiveness of MR for performing scientific tasks against in-person and online meetings. We will then evaluate how different team compositions of human users and AI agents working in MR could enhance or hinder their scientific tasks. The proposed project has the potential to enable the metaverse as a tool for collaboration that can transform teamwork, creativity, and human-AI collaborations.

AI Innovation: To answer our research questions, we are conducting a 2 (use of AI: AI vs no-AI) x 3 (medium: in-person, Zoom, MR) behavioral experiment in which participants must collaborate on two creativity tasks. While participants are randomly assigned participants to the medium, they will complete one creativity task using AI, and another one without it. This experimental design involves randomizing the order of the AI-presence condition and randomly assigning participants into mediums. In short, here are the descriptions of the experimental conditions: In-person: Participants work on the tasks in the same room. When interacting with AI, participants will have a tablet where they will work with an AI agent using their voices. Zoom: Participants will work on tasks in separate rooms. Each room has a computer with Zoom where participants work together. When interacting with AI, participants will see a new participant representing the AI agent. Mixed Reality (MR): Participants will work on the task in separate rooms. Then, participants will use MR headsets to communicate and complete the task. They will see each other as avatars interacting in the same room. Participants will create the avatars before running the session. When interacting with AI, participants see a new avatar that embodies the AI agent. We have already deployed a multi-player Mixed Reality (MR) condition using Unity, a game-engine platform often used to create VR/MR applications. To enable multi-player interactions, we implemented our application using an

open-source framework called Photon, which allows multiple users to connect and visualize changes in their environment. We integrated our application with OpenAI's RealTime API to develop an AI agent that interacts with team members. This API allows sending users' voices directly to the API, and returns the voice generated by a Large Language Model. This configuration allows very fast communication between team members and AI. We piloted this MR application with 30 participants to improve the application and learn from the differences between MR and other applications.

Our preliminary results show that MR can support social relationships better than Zoom. Moreover, MR participants reported communications that were more similar to in-person communication than Zoom participants did. Moreover, participants reported that MR enabled them to be together socially while being in different places. Some participants reported that the environment allowed them to be more creative and connect better with their teammates. These results suggest that MR can increase social presence and improve teams' social cohesion. Despite these positive results, we found that MR participants produced fewer ideas than in-person or Zoom participants. These findings, however, were highly mediated by how supportive participants found working in an MR environment.

Translational Evidence: This research demonstrates translational potential by illustrating how AI-integrated MR technologies can significantly enhance collaboration, creativity, and productivity in remote or hybrid work environments. By enabling interactions between human users and AI agents within immersive virtual spaces, the developed AI solution could profoundly transform the interactions between AI and humans by providing space, movement, and embodiment. This approach is positioned to significantly impact socio-economic development by expanding access to high-quality collaboration opportunities beyond geographic and resource limitations, enabling more inclusive and effective global research initiatives, knowledge sharing, and innovation ecosystems.

RISE Dimensions: The innovation aligns strongly with multiple RISE dimensions, notably Inclusion, Responsibility, and Ethics. Our metaverse solution supports Inclusion by removing barriers to effective collaboration, thus enabling diverse participation across geographical, cultural, or economic boundaries. Our work considers Responsibility through its intentional design, explicitly considering the impacts of human-AI collaboration on creativity, productivity, and social interaction within scientific teams. Lastly, Ethical alignment is important through deliberate experimentation assessing the integration of AI agents and ensuring their interactions promote beneficial outcomes while mitigating risks of dependency or negative social dynamics. Collectively, these dimensions illustrate the thoughtful embedding of RISE values into the development and deployment of AI-powered MR technologies for the broader social good.

References:

[1] Ball, M. 2022. The Metaverse: And How It Will Revolutionize Everything. Liveright Publishing Corporation, a division of W.W. Norton.

[2] Gómez-Zará, D. et al. 2023. The promise and pitfalls of the metaverse for science. Nature human behaviour. 7, 8 (Aug. 2023), 1237–1240.

3] Li, X. et al. 2018. A critical review of virtual and augmented reality (VR/AR) applications in construction safety. Automation in construction. 86, (Feb. 2018), 150–162.

[4] Torro, O. and Pirkkalainen, H. 2023. Design principles for social exchange in social virtual reality-enabled virtual teams. Virtual reality. 27, 4 (Dec. 2023), 2791–2820.

#81: A Checklist for Trustworthy, Safe, and User-Friendly Mental Health Chatbots

Authors: Shreya Haran, Samiha Thatikonda, Dong Whi Yoo and Koustuv Saha

Mental health concerns are rising globally, prompting increased reliance on technology to address the demand-supply gap in mental health services. In particular, mental health chatbots are emerging as promising solution, but remain largely untested, raising concerns about safety and potential harms. In this paper, we dive into the literature to identify critical gaps in the design and implementation of mental health chatbots. We propose a preliminary checklist to help guide the development and design of more trustworthy, safe, and user-friendly chatbots. The checklist serves as both a developmental framework and an auditing tool to ensure ethical and effective chatbot design. We discuss how implementing such a checklist can enhance self-management support for mental health by reducing potential risks while also inspiring the development, refinement, and adoption of similar frameworks as a standard for ethical and effective technology use in sensitive contexts.

#107: AI and the Eucharist: Practical Analysis and Theologically-Informed Ethical Reflections

Authors: Megan Angell

With the rise of ChatGPT and other public-facing large language models (LLMs), people have increasingly turned to them for help with a variety of tasks, from mundane work emails to travel planning to coding. This research explores the usefulness, responsibility and ethicality of large language models at helping individuals with a far more personal topic: their religious faith. I analyze two leading general purpose LLMs and two religiously-focused LLMs to evaluate their ability to explain a particularly complex but foundational theological principle, that is, the Catholic belief of the Eucharist. In particular, I examine whether LLMs can address a common pattern of misunderstandings about the Eucharist that I identify from national survey data on American Catholics' religious beliefs. Finally, I analyze the ethical implications of learning about intimate religious topics from an AI and whether reflecting on AI itself may further the Catholic Church's understanding of its own doctrine.

First, this interdisciplinary research provides an extended analysis of survey data on American Catholics' beliefs about the Eucharist, the sacrament that Catholics believe is the real presence of Jesus. In recent years, three major recent surveys on the topic were conducted by secular and religious institutions and reveal widespread misunderstanding of Catholic doctrine. However, these datasets remain under-studied, especially with regard to respondents' race/ethnicity and socioeconomic status. I draw from consumer segmentation practices to analyze the data and develop typologies of respondents on the basis of religious involvement, faith knowledge, and socioeconomic demographics. Using these typologies, I assess the religious educational needs of each group with particular attention to diverse and underserved communities.

Second, this paper explores the suitability of using LLM tools to provide religious catechetical education which have the potential to be especially valuable for underserved populations who may not have access to traditional means of religious education such as specialized publications and in-person programming in their preferred language. For example, the widely popular, Catholic app Hallow has incorporated a specialized LLM, MagisteriumAI into its free version to provide religious education in a wide variety of native languages. However, there is little independent inquiry about whether this is a responsible use of AI. For example, the Eucharist is a complex theological belief at the core of the Catholic faith and AI "hallucinations" or other inaccuracies have the potential to unintentionally lead users' religious faith astray. To address this concern, I analyze the ability of 2 secular LLMs, ChatGPT 4.0 and DeepSeek Deep Reasoning, and 2 Catholic LLMs, MagisteriumAI and CateGPT, to educate about the Catholic faith.

Finally, this paper explores the ethical implications of learning about an intimate topic from an AI and how AI itself may prompt theological reflection among Church experts. What does it mean to learn about God's presence in the Eucharist from an entity that is definitionally absent but appears to be human-like? What does it mean, especially for vulnerable populations, to learn about the Eucharist from an AI which cannot "know" vulnerability in any real sense, and which is not embodied in a gender? These ethical reflections draw from academic data science literature on AI explainability and manipulation and the diverse Catholic philosophical tradition.

#111: Co CoT: A Prompt-Based Framework for Collaborative Chain-of-Thought Reasoning

Authors: Seunghyun Yoo

Due to the proliferation of short-form content and the rapid adoption of AI, opportunities for deep, reflective thinking have significantly diminished—undermining users' critical thinking and reducing engagement with the reasoning behind AI-generated outputs. To address this issue, we propose an Interactive Chain-of-Thought (CoT) Framework that enhances human-centered explainability and responsible AI usage by making the model's inference process transparent, modular, and user-editable. The framework decomposes reasoning into clearly defined blocks that users can inspect, modify, and re-execute, encouraging active cognitive engagement rather than passive consumption. It further integrates a lightweight edit-adaptation mechanism inspired by preference learning, allowing the system to align with diverse cognitive styles and user intentions. Ethical transparency is ensured through explicit metadata disclosure, built-in bias checkpoint functionality, and privacy-preserving safeguards. This work outlines the design principles and architecture necessary to promote critical engagement, responsible interaction, and inclusive adaptation in AI systems aimed at addressing complex societal challenges.

#116: Human-Centered AI for Early Detection and Prevention of Carjackings in Underserved Urban Highways

Authors: Tímar Contreras, Matias Soto and Jorge Vásquez

Abstract. Carjackings—locally referred to as encerronas—are a growing threat in major Latin American urban highways, particularly in Chile. These crimes typically involve armed attackers targeting vehicles in motion, often in high-speed or isolated road segments. Victims face significant physical and psychological risks. Underserved communities are especially vulnerable due to limited surveillance coverage, scarce police presence, and minimal access to early-warning technologies. Traditional safety systems are reactive, expensive, and not predictive. We propose a humancentered, low-cost, explainable AI system for early detection and prevention of carjackings in highway settings. Our system processes short video clips (2–5 minutes) recorded by patrol units or civilian inspectors. It integrates a hybrid AI architecture consisting of: (i) YOLOv8 for real-time object detection to identify vehicles, road elements, and spatial layouts; and (ii) lightweight temporal transformers to model suspicious behavioral patterns, such as stop-and-go traffic, vehicle clustering, and lane blocking. As a second phase, the system incorporates scene understanding through human keypoint detection, enabling the analysis of body posture, motion intent, and group coordination of individuals involved in potential carjackings. This fine-grained layer provides additional context to differentiate threatening human actions (e.g., approaching aggressively, coordinated blocking) from benign activity. The training set includes real and synthetically generated highway scenes enriched with contextual metadata (road topology, lighting conditions, and known crime zones). Human experts review and validate outputs to minimize false positives and bias. We deploy this solution as a Software-as-a-Service (SaaS) platform optimized for low-bandwidth and constrained hardware settings. Users upload video footage and receive:

- Georeferenced heatmaps indicating risk hotspots,
- Anomaly summaries classifying type and severity,
- Actionable reports supporting patrol planning and decision-making.

Additionally, we are developing a mobile and desktop application to provide real-time alerts directly to highway operators. This app enables centralized monitoring of multiple highway cameras and facilitates seamless integration with law enforcement agencies, improving response times and situational awareness during active threats.

Preliminary Results and Limitations. Initial evaluations were conducted on a custom dataset of 124 annotated highway videos (totaling 14.3 hours of footage) representing a mix of normal traffic and simulated threat scenarios. The system achieved a precision of 91.7%, a recall of 87.9%, and a false positive rate below 3.2%. Using a Jetson Nano edge device, the average inference time was 640 ms per video segment. User testing with 12 traffic inspectors and 3 law enforcement analysts showed positive feedback on system usability, alert clarity, and report utility. Identified limitations include reduced accuracy in low-light or high-glare conditions, and challenges in distinguishing between abnormal congestion and deliberate obstruction. Ongoing work focuses on real-time feed integration, adaptive thresholding, and

broader anomaly classification. **RISE Framework.** The system adheres to Responsible, Inclusive, Safe, and Ethical (RISE) AI principles: Responsibility: AI-generated alerts are always subject to human validation, with interpretable confidence scoring and auditability. Inclusion: Designed for low-resource settings, it promotes equitable access to AI-enhanced public safety. Safety: The system prioritizes early detection to protect civilians before crimes escalate. Ethics: Feedback from affected communities and diverse data sources helps mitigate bias and ensures relevance.

Interdisciplinary Contribution. This project bridges AI, urban policy, public safety, and human-centered design. It fosters collaboration among academic researchers, highway operators, law enforcement agencies, and local communities. By addressing a pressing real-world issue through scalable technology, it contributes to the global discourse on ethical, real-time AI for vulnerable populations.

Conclusion. This work showcases the potential of targeted, explainable, and inclusive AI to democratize access to public safety technologies. With further development, it can inform the design of early-warning systems for a broader range of threats, paving the way for safer and more equitable urban mobility.

#145: A Personalized, AI-Driven Visual Intervention for Lifelong Emotional Well-being

Authors: Agnieszka Marczak-Czajka, Michael Villano and Adam Czajka

Recent advances in Artificial Intelligence and affective neuroscience are enabling novel approaches to understanding and influencing human emotions through AI-driven visual and auditory stimulations. One emerging area involves the use of AI-generated visual stimuli to modulate affective states. Preliminary findings, such as those reported by Marczak-Czajka et al. (Marczak-Czajka A, Redgrave T, Mitcheff M, Villano M, Czajka A., "Assessment of human emotional reactions to visual stimuli 'deep-dreamed' by artificial neural networks," Frontiers in Psychology, Dec. 2024) indicate that imagery synthesized to maximize activations in selected fractions of deep neural networks (e.g., techniques like "deep dream") may elicit varying emotional responses, which are hypothesized to stimulate selected brain sections more precisely than naturalistic images. While these results are exploratory in nature, they open promising directions for integrating machine-generated semantics-free stimuli to offer human-centered emotional regulation strategies, and – in the long term – affordable mental healing approaches. This paper proposes a generalizable, adaptive mechanism for promoting emotional well-being that leverages AI-generated visual stimuli and real-time eye tracking put in a feedback loop. The proposed system operates by capturing users' visual attention patterns (including gaze fixation, saccades, and pupil dilation) as implicit indicators of cognitive engagement and affective state. These signals, which can be collected passively through a consumer-grade eye-tracking technology (e.g., via webcams or mobile cameras), are then used to dynamically tailor the visual input in ways that align with a target emotional trajectory. For example, individuals experiencing stress might be presented with synthesized visual patterns known to reduce arousal, while those facing fatigue could receive synthesized imagery calibrated to stimulate attentional re-engagement. The authors will share with the audience their preliminary quantitative results and observations after running two studies with 100+ online users reacting to synthetically generated ("deep dreamed") visual stimuli, and 50+ stationary users (Notre Dame students), reacting to synthetically-generated images presented in a Virtual Reality environment. The proposed mechanism very well aligns with the United Nations Sustainable Development Goal: ensuring healthy lives and promoting well-being for all at all ages by offering scalable, customizable, and ethically sound support for emotional health across diverse life stages and environments. The proposed approach is conceived as a platform-agnostic and age-inclusive intervention that can function across different contexts, from mental wellness tools and therapeutic support systems to educational tools. It may offer particular value in settings where verbal self-report is unreliable or impractical, such as with young children, elderly populations, subjects suffering from brain degeneration (such as dementia or Alzheimer disease), or individuals with communication challenges. Moreover, its closed-loop design, that is, integrating stimulus generation with continuous, adaptive feedback, may offer a degree of personalization not typically available in standard wellness interventions. The proposed mechanism lays the groundwork for a new class of emotional well-being enhancement tools, but may be also research-enabling: through aggregated, anonymized data collection, it may support the long-term investigation of how computational visual features interact with human affective and attentional processes across populations.

#160: Conversation Dynamics in Human-AI Collaborative Teams: Study of Group Decision-Making in Mixed-Agent Environments

Authors: Jiehui Luo and Ahmed Abbasi

As generative AI systems increasingly integrate into collaborative work environments, we face the critical challenge of understanding how these systems influence team dynamics, particularly in complex decision-making tasks where human judgment remains essential. The integration of AI into group settings introduces unique social and cognitive dynamics that can significantly impact team performance, trust relationships, and decision quality. Current research has predominantly focused on one-to-one human-AI interactions, leaving substantial gaps in our understanding of how these dynamics scale to group settings with multiple human collaborators.

Our research focuses on an experimental framework for studying human-AI collaboration in small groups (triads) performing complex evaluation tasks such as essay grading. We systematically vary AI interaction modalities, visibility, and agency within these groups to examine how these factors influence conversation patterns, trust development, and collaborative outcomes. By applying natural language processing techniques to analyze conversational data and combining this with psychometric measurements, we quantify how AI integration affects social dynamics including confidence expression, trust formation, and authority distribution within mixed-agent teams.

Our preliminary findings demonstrate that the mere presence of AI in group settings significantly reshapes conversation dynamics, with both beneficial and problematic outcomes. In pure generative AI scenarios, we observed a notable alignment phenomenon where multiple AI agents converged on similar judgments when confronted with cases where human evaluators showed high variance in their assessments. This "AI consensus effect" appears to exert social influence on human decision-makers, potentially reducing the diversity of perspectives in mixed teams. Additionally, we found that the cognitive processing level of conversations evolved significantly across discussion rounds, with AI contributions initially elevating analytical depth but gradually shifting toward simpler consensus-building exchanges in later rounds. Our ongoing research will expand these investigations to examine various human involvement configurations, testing how expertise distribution, role clarity, and team composition influence critical evaluation processes and deference to AI judgments. These insights aim to provide actionable guidance for designing AI systems that can be effectively deployed in educational assessment contexts and other collaborative environments where preserving the diversity and depth of human judgment remains crucial.

This work addresses multiple RISE dimensions, particularly responsibility and inclusion. From a responsibility perspective, we identify specific conversation patterns that indicate when teams are overly deferring to AI judgment versus maintaining appropriate critical evaluation. Our inclusion focus examines how different AI integration approaches can either amplify or mitigate existing power imbalances in group settings, particularly for team members with less domain expertise or social confidence. By developing empirically-grounded frameworks for how AI systems should present information and uncertainty in group contexts, we contribute to more ethical design principles for collaborative AI systems that support rather than supplant human decision-making capabilities.

#163: RISE Together: Empowering Citizens Through Ethical AI Navigation of County Services

Authors: Keyang Zhou, Rachel DeGaugh, Jing Peng and Si Chen

County government websites serve as essential gateways to public services, yet their fragmented navigation structures create substantial barriers for diverse resident populations. This challenge is particularly pronounced for international students and temporary residents who must navigate multi-jurisdictional policies spanning federal immigration law, local county services, and institutional requirements while maintaining legal compliance with their visa status. We developed an AI-powered navigation system that addresses these complexities through a five-layer architecture encompassing user interaction, inference processing, specialized language models, real-time data extraction, and privacy protection. Our approach employs a multi-agent framework where specialized AI components handle immigration law, work authorization, local benefits, and compliance verification to deliver status-aware guidance. The system fills critical gaps in existing government tools by providing visa-specific eligibility filtering, legal compliance verification, and coordinated recommendations across multiple agencies. For international students, this enables safe navigation of available services while preventing inadvertent visa violations that could jeopardize their legal status. Current implementation includes a completed frontend interface and database infrastructure, with AI model integration in progress for full system functionality. We evaluate performance through comparative benchmarking against existing AI assistants using realistic navigation scenarios, measuring task completion accuracy, usability metrics, and safety compliance. This research demonstrates how responsible AI design can bridge complex policy environments to provide equitable service access for underserved populations.

#167: GG or Rage Quit? Combating Toxicity in Competitive Online Gaming and Examining Peer Influence Across Sequential Matches

Authors: Sunan Qian, Corey Angst and Yoonseock Son

Multiplayer online battle arena (MOBA) games such as League of Legends (LoL) and Defense of the Ancients (DotA) possess millions of users globally (Wong, 2025). However, toxic behavior among players, which leads to unpleasant experiences and harms community engagement, has become a significant concern in online gaming, especially for MOBAs like Dota 2 that are more competitive (Talbot, 2019). Gaming companies have been striving to address toxicity in different ways. A coalition including over 30 major gaming companies such as Blizzard, Riot, CCP, Twitch, Discord, and Epic has formed the Fair Play Alliance, which aims to share research and best practices to better understand toxicity (Moore, 2018). Ubisoft has co-developed the "Good Game Playbook," which is a guide distributed to players who report or are reported for disruptive behavior (O'Connor, 2023). Electronic Arts (EA) has introduced a "Positive Play" charter which lays down clear guidelines and expectations for player behavior. (Hetfeld, 2020). Therefore, our research aligns with the goal of the gaming companies to better understand and mitigate toxic behavior, in order to maintain a positive gaming environment.

The notion of toxicity is often composed of a range of actions, including disruptive gameplay like cheating, griefing and spamming, as well as abusive communications in the form of harassment, flaming and especially verbal abuse (Adinolf & Turkay, 2018; Beres et al., 2021; Foo & Koivisto, 2004; Frommel et al., 2023; Kwak & Blackburn, 2014; Lapidot-Lefler & Barak, 2012; Neto et al., 2017; Shen et al., 2020; Türkay et al., 2020). Verbal abuse stands out as one of the most prominent and impactful forms of toxic communication, which makes it the central focus of this research. In this study, we primarily focus on the textual toxicity in chat messages, and we intend to explore three research questions: 1) How can we accurately label toxic chat within each match? 2) How does exposure to toxicity in previous matches influence a focal player's toxicity in future matches? 3) How does the contagion effect change as time goes by?

Our data comes from the popular game of Dota 2 that was released by Valve Corporation in 2013 as a continuation to DotA. We randomly select 1,528 players who tend to play in an English-speaking region and retrieve information for their consecutive matches, including texual chat history, match outcome, and performance metrics, etc. Our textual chat history data consists of all utterances in the public channel, where all 10 players from both teams can contribute to the conversation. Our final sample consists of 1,358 players tracked over 30 days from December 2024 to January 2025, which gives us a total of 51,771 player-match observations.

We measure the toxicity of each utterance by applying transfer learning and fine-tuning a bi-directional encoder representations from transformers (BERT)-base model for multi-class toxicity classification on an annotated Dota 2 chat message dataset called CONDA (Weld et al., 2021). The CONDA training set contains 26,921 utterances from 1,911 matches. The CONDA validation and test sets each contain 8,974 utterances from 1,778 matches. Each utterance belongs to exactly one of the four classes: explicit toxicity (E), implicit toxicity (I), action (A), and others (O). Explicitly toxic utterance usually contains toxic words with the intent to insult or humiliate others. Implicitly toxic utterance refers to hidden toxicity that usually cannot be seen from the text itself, including sarcasm inferred from context. We fine-tune our BERT-base model using a grid search over different sets of hyperparameters, and we obtain a mean AUC score of 0.96 across all combinations. We preprocess the data strictly following the procedure in Weld et al. (2021) to construct our own test set. One crucial step in preprocessing is to merge consecutive utterances by a single player into one, with a special token "[SEPA]" in between each original utterance. By doing so, we effectively consider the context of each conversation. Context is crucial in gaming, because a seemingly toxic word, such as curse language, can express excitement without encompassing any toxic intention; on the other hand, seemingly harmless language can also very well be sarcastic under certain context (e.g., "EZ").

The explicit and implicit toxicity scores for each utterance are measured as the probability of this utterance belonging to class E and I, respectively. Then we sum up the explicit (implicit) toxicity scores for all utterances by each player in each match, in order to form an aggregate explicit (implicit) toxicity score as our dependent variables. Since we would like to study the effect of exposure to toxicity, we sum up the lagged explicit (implicit) toxicity scores of focal player's teammates as our first set of independent variables. Then our second set of independent variables is defined by summing up the lagged explicit (implicit) toxicity scores of focal player's opponents. We separate teammate versus opponent toxicity, because in a public channel, the toxicity from teammates is very likely to be directed towards the opposing team. Our moderator is the time interval in minutes between two consecutive matches. To study our research questions, we first utilize the Allison (2009) hybrid model to decompose within- and between-player effects. This approach is especially helpful because it allows us to distinguish stable, between-player differences (e.g., some players consistently experience higher levels of toxicity across matches) from dynamic, within-player fluctuations (e.g., how a player's behavior changes following specific toxicity exposure).

Results show statistically significant between-player effects for both explicit and implicit toxicity, but minimal to none within-player effects, which suggests that stable differences between people matter more than short-term fluctuations within a person. Theoretically, this implies that players may not respond uniformly to toxic environments. Therefore, we are motivated to adopt a k-means clustering approach and partition our players into four clusters. Moreover, to address potential endogeneity, we also apply a two stage residual inclusion (2SRI) framework with instrumental variables (Angrist & Pischke, 2009; Cameron & Trivedi, 2010; Kennedy, 2008). Our instruments are the lagged number of words produced by teammates (opponents) in the previous match, and are strongly related to the independent variables according to the partial F-stat. Theoretical logic also supports its validity. More teammate or opponent chat naturally increases the chance for those players to produce toxic content, which makes word count relevant to their subsequent toxicity. At the same time, our instruments are exogenous and only influence the dependent variables through the independent variables – there is no direct psychological pathway through which other players' prior word volume would influence focal player's toxicity in the next match.

Overall, results show a statistically significant and positive effect of lagged opponent toxicity on focal player toxicity, which suggests that contagion is present, with toxicity from the opposing team spreading across matches. Moreover, players with significantly higher baseline toxicity are less sensitive to opponent toxicity. In addition, as time goes by, focal player toxicity decays. However, interval between matches does not influence the relationship between lagged teammate or opponent toxicity on focal player toxicity. For gaming companies, one implication of the findings is that interventions like toxicity warnings or reporting systems may not work uniformly across users. Gaming companies could implement tiered interventions: empathetic nudges or feedback for low-to-medium toxic players, versus stronger penalties or gatekeeping mechanisms for habitual offenders. Moreover, simply giving players time between matches won't block the contagion of toxicity, which suggests that passive time delay is insufficient – active behavioral nudges, toxicity-aware matchmaking, or temporary role restrictions are required to block the transfer of toxic norms.

By leveraging a novel AI-driven econometric approach, the study contributes to the literature by combining instrumental variable two-stage residual inclusion (2SRI) modeling with k-means clustering to disentangle the dynamic, heterogeneous effects of exposure to teammate and opponent toxicity on subsequent player behavior. This methodological innovation enables causal inference in the presence of endogeneity and unobserved player heterogeneity, which offers a more rigorous understanding of behavioral contagion in gaming environments. Insights from this study can inform platform-level interventions with strong policy relevance, such as dynamic matchmaking and targeted behavioral nudges. Aligned with the RISE dimensions, this research promotes Responsibility through precise machine learning and econometric modeling, supports Inclusion by identifying subpopulations at risk of marginalization due to toxic play, enhances Safety by informing preemptive design strategies, and upholds Ethics by enabling more equitable and evidence-based governance of online communities.

Reference:

- Adinolf, S., & Turkay, S. (2018). Toxic behaviors in Esports games: player perceptions and coping strategies. Proceedings of the 2018 Annual Symposium on computer-human interaction in play companion extended abstracts.
- Allison, P. D. (2009). Fixed effects regression models. SAGE publications. Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.
- Beres, N. A., Frommel, J., Reid, E., Mandryk, R. L., & Klarkowski, M. (2021). Don't you know that you're toxic: Normalization of toxicity in online gaming. Proceedings of the 2021 CHI conference on human factors in computing systems.
- Cameron, A. C., & Trivedi, P. K. (2010). Micro-econometrics using Stata (Vol. 2). Stata press College Station, TX.
- Foo, C. Y., & Koivisto, E. M. (2004). Defining grief play in MMORPGs: player and developer perceptions. Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology.
- Frommel, J., Johnson, D., & Mandryk, R. L. (2023). How perceived toxicity of gaming communities is associated with social capital, satisfaction of relatedness, and loneliness. Computers in Human Behavior Reports, 10, 100302.
- Hetfeld, M. (2020, June 18). EA promises increased measures to combat toxicity in its games. *PC Gamer*. Link
- Kennedy, P. (2008). A guide to econometrics. John Wiley & Sons.
- Kwak, H., & Blackburn, J. (2014). Linguistic analysis of toxic behavior in an online video game. International conference on social informatics
- Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. Computers in human behavior, 28(2), 434-443.
- Moore, B. (2018, March 22). Major game companies are teaming up to combat toxicity in gaming. *PC Gamer*. Link
- Neto, J. A., Yokoyama, K. M., & Becker, K. (2017). Studying toxic behavior influence and player chat in an online video game. Proceedings of the international conference on web intelligence,
- O'Connor, D. (2023). Combating Toxicity With The Good Game Playbook. *Ubisoft*. Retrieved April 12. Link
- Shen, C., Sun, Q., Kim, T., Wolff, G., Ratan, R., & Williams, D. (2020). Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. Computers in human behavior, 108, 106343.
- Talbot, C. (2019). According to a new study Dota 2 has the most toxic online community. *PCGamesN*. Retrieved April 12 from Link
- Türkay, S., Formosa, J., Adinolf, S., Cuthbert, R., & Altizer, R. (2020). See no evil, hear no evil, speak no evil: How collegiate players define, experience and cope with toxicity. Proceedings of the 2020 CHI conference on human factors in computing systems
- Weld, H., Huang, G., Lee, J., Zhang, T., Wang, K., Guo, X., Long, S., Poon, J., & Han, S. C. (2021). CONDA: a CONtextual Dual-Annotated dataset for in-game toxicity understanding and detection. arXiv preprint arXiv:2106.06213.
- Wong, M. (2025). Dota 2 Player Count: How Many People Play The Game in 2025? esports.net. Retrieved April 12 from Link

#168: Detecting and Mitigating Risk in Child–LLM Interactions: A Human-Centered Framework for Generative AI Safety

Authors: Yaman Yu

As children increasingly engage with large language models (LLMs) through conversational platforms, new safety challenges are emerging that current AI safeguards fail to address. These systems are not designed to account for the developmental, emotional, and relational needs of children. Through prolonged interactions, children may encounter risks such as emotional manipulation, dependency formation, and inappropriate roleplay dynamics. These harms often evolve gradually and remain undetected by existing moderation mechanisms that are optimized for single-turn prompts and adult users. This project introduces YouthSafe, a human-centered framework for understanding and mitigating risks in child–LLM interactions. We developed a secure and IRB-approved infrastructure for collecting anonymized real-world chat histories from teens using generative AI systems. Based on this dataset, we constructed a multi-level risk taxonomy focused on developmental vulnerabilities, including cognitive distortion and relational boundary violations. Building on these insights, we designed and prototyped an intervention model that applies multi-turn context reasoning to detect risky trajectories and generate actionable responses. The system supports both direct mitigation during interaction and caregiver-informed interventions. By combining empirical data analysis, youth-centered design, and responsible AI methods, this work redefines AI safety for child users as an issue of longitudinal interaction risk rather than content filtering alone. Our contributions inform future research and policy focused on building generative AI systems that are safe, context-aware, and developmentally aligned.

#17: FAULT: Failure Analysis Using Learning Techniques for soft actuators

Authors: Rishabh Goel, Tejonidhi Deshpande, Tingyu Cheng and Josiah Hester

This abstract was rescheduled to present a poster in the session on October 8, 2025.

Abstract: Soft actuators, particularly those of bio-hybrid and biodegradable nature, confront formidable challenges in natural settings, often succumbing to failures due to structural material degradation. These kinds of actuators hold immense potential for environmentally friendly and sustainable applications, but their real-life implementation is hindered by the rapid degradation of their structural material. Recognizing the need for an intelligent failure detection system, we propose a novel scheme for soft actuators that extracts relevant features from time series data and employs different classifiers for classification. This system not only identifies the location of the failure on the robot but also provides a failure severity score, potentially enabling precise adjustments to the actuator controller based on its current state. We also developed a multi-chamber pneumatic soft bending actuator that can achieve different bending configurations depending on how much each chamber is inflated. To make development fast and efficient, we train and test our failure detection scheme on multi-chamber actuators made of silicone as a proxy for robots made from bio-hybrid/bio-degradable materials such as gelatine and hydrogel. Finally, we present a comparison between model accuracy of different classifiers and achieve up to 98% accuracy in classifying the location and severity of a failure across 5 different prototypes made from 2 different materials.

Wednesday, October 8, 2025:

Track: AI for Safety and Safety of AI

#32: Playing Fair in Dire Straits: Bias Reduction for Humanitarian AI

Authors: Cameron Kormylo and Marialena Bevilacqua

Problem Statement: Humanitarian organizations are racing to harness artificial intelligence (AI) to revolutionize disaster response and save more lives. One notable use case is the use of AI to predict the severity of disasters and assess their impacts. For example, AI models can rapidly analyze satellite imagery to assess infrastructure damage - e.g., flooded areas, downed buildings - allowing rescuers to gauge a disaster's toll and allocate resources effectively. Notably, McKinsey's 'Noble Intelligence' initiative slashed building damage assessment time from weeks to minutes by processing multi-source data, guiding relief planners to prioritize hardest-hit areas (Heteren et al., 2020).

However, the data used to train these models are often sourced from online crowdsourcing platforms, where workers quickly and cheaply label datasets. This introduces a critical vulnerability: biases from human labelers can infiltrate the data, causing AI systems to learn skewed "ground truths," further undermining fairness. Literature confirms that labelers' backgrounds and personal opinions significantly influence judgements (Eickhoff 2019). Importantly, biases in humanitarian contexts may have particularly severe consequences, such as misdirecting life-saving resources, disproportionately harming underserved communities, and deepening inequalities. These inequalities could arise from a labeler's innate tendency to overestimate disaster severity for populations demographically or geographically closer to themselves. Beyond a fundamental technical flaw, such occurrences introduce moral compromise to humanitarian missions, where impartiality is a bedrock principle.

Previous efforts have explored various interventions to reduce labeling bias, including worker selection and training (Barbosa & Chen 2019), task design (Hube et al., 2019), and post-hoc statistical bias mitigations (Kamar et al., 2015). However, these interventions may not apply in humanitarian contexts, where ethical principles like equity and fairness take precedence. Humanitarian biases often stem from deep societal factors such as historical inequities and cultural stigmas (Tremblay-Boire & Prakash 2019), making general interventions inadequate. Additionally, the urgent nature of humanitarian crises limits the time and resources available for implementing sophisticated bias-mitigation solutions.

AI Innovation: Increasingly, literature has supported the role of "Games with a Purpose" (GWAPs) in achieving meaningful outcomes by harnessing human intelligence for solving complex tasks (Von Ahn & Dabbish, 2008). They strategically integrate enjoyable gameplay with valuable data collection, human computation, or problem-solving tasks that are traditionally difficult to accomplish. Importantly, GWAPs have demonstrated their ability to accomplish these outcomes without sacrificing efficiency. This is exemplified by the classic ESP Game, which famously collected over 1.2 million image labels from approximately 14,000 players in just one month, maintaining high engagement thereafter (Von Ahn & Dabbish, 2004).

cBuilding on this approach, we propose an innovative game designed to reduce labeler bias in humanitarian AI training. Our two-phase study focuses on labeling disaster aftermath images for severity. Phase 1 provides baseline findings to inform the design of our game by assessing the severity of "distance"-based labeler bias (e.g., how do labels change when labelers identify context cues signaling the geographic location of the disaster or the demographic details of affected populations?). In Phase 2, we design a GWAP intended to mitigate these biases, where players act as disaster response coordinators, labeling images to effectively allocate aid. Subtle mechanics - like intermittent "bias-check prompts" (e.g., revising earlier images with altered contexts to test for uniform judgement) - are woven into gameplay, steering labelers toward impartial judgements without overt correction. Consistency scores are calculated based on how uniformly players rate similar images regardless of contextual cues, nudging them towards unbiased patterns. This preemptive, behavioral approach ensures engagement and efficiency while producing fairer data for AI models guiding humanitarian aid.

Translational Evidence: Our gamified labeling approach offers humanitarian organizations a scalable, cost-effective tool for improving AI-driven disaster response. Organizations like the Red Cross or Catholic Relief Services could integrate this method into existing crowdsourced workflows. Beyond humanitarian contexts, this approach could benefit sectors like recruiting (e.g., labeling job applicants) or social media (e.g., identifying hate speech), where unbiased crowdsourced data is critical for equitable outcomes. The design's flexibility - customizable to different labeling tasks by adjusting in-game assets, prompts, and scoring mechanics - broadens its applicability across domains. Socio-economic impacts include more equitable aid distribution, particularly benefiting underserved communities in remote or demographically distinct regions (e.g., rural areas, minority populations) that biased models might otherwise overlook. By amplifying the quality of crowdsourced data, this innovation could save lives and reduce recovery disparities, aligning with humanitarian goals of impartiality and justice.

RISE Dimensions: Our approach emphasizes multiple RISE dimensions. Responsibility is at the core of our proactive effort, ensuring AI serves humanitarian missions justly by mitigating bias at the source. By prioritizing data integrity, we uphold accountability to affected populations. Our focus on reducing "distance"-based bias amplifies infusion and ensures the needs of underrepresented communities are not discounted due to labeler prejudice. Similarly, ethics underpins the design, as we address the moral imperative of impartiality in life-saving contexts. Drawing from computer science, social sciences, and ethics, this interdisciplinary approach exemplifies how responsible, inclusive, and ethical AI can reshape disaster response through a continued focus on societal good.

#49: A landscape analysis of AI transparency standards and opportunities for enhancement

Authors: Alexis Baria

AI standards serve as tools to translate ethical and responsible AI principles into practical guidelines and processes. A common purpose among these standards is transparency, or the general goal of providing greater understanding of an AI system's origin, presence, and operations. This paper provides a comprehensive overview of existing transparency standards, focusing on their varying objectives, scopes, and definitions. Our analysis has three primary aims: outlining the current landscape of AI transparency standards, examining the challenges associated with formalizing differing transparency values and perspectives, and identifying opportunities for enhancement. We found commonalities across standards including a desire for applicability across various technologies and usage domains, as well as tiered transparency levels designed for distinct stakeholder needs. We also found challenges regarding implementation, such as assessing risk to indicate what level of transparency is needed, and determining what specific methods providers should use. Moreover, our review indicates that standards differ in their conceptualization of transparency, viewing it mostly as an intrinsically valuable attribute or in terms of its relational impact on stakeholders. Finally, this paper will highlight key areas for improvement and will aim to contribute to ongoing efforts towards developing transparency standards that effectively address the diverse needs of stakeholders.

#58: SAFES: Sequential Privacy and Fairness Enhancing Data Synthesis for Responsible AI

Authors: Spencer Giddens and Fang Liu

Problem Statement: Data-driven and AI-based decision-making is being increasingly adopted in many disciplines, and data collected as part of this process often contain sensitive information about individuals. These data are frequently used to make socially impactful decisions including, but not limited to, loan approvals, hiring decisions, and criminal justice decisions. Although such applications can and do have legitimate societal benefits, it is of paramount importance to ensure that the use of such sensitive data is responsible and carried out with the highest possible ethical standards.

Privacy and fairness are two important ethical concerns when working with sensitive personal data to train machine learning (ML) and AI algorithms. Although differential privacy (DP) provides a robust framework for guaranteeing privacy and several widely adopted methods have been proposed for improving fairness, the vast majority of existing literature treats the two concerns independently. This is especially concerning since previous research has shown that the application of DP tends to magnify unfairness, and fairness transformations likewise unevenly distribute privacy risk to underprivileged groups. For methods that consider privacy and fairness simultaneously, they often only apply to a

specific machine learning task, limiting their generalizability. Since datasets with privacy concerns are likely to also have fairness concerns and vice versa, it is critical to develop efficient privacy- and fairness-enhancing methods for releasing and analyzing data.

AI Innovation: There remains a scarcity of research on general-use synthetic data that both satisfies DP guarantees and reduces unfairness, representing a critical area for advancing responsible AI. We address this gap by proposing SAFES -- a Sequential privacy and Fairness Enhancing data Synthesis procedure -- which sequentially combines DP data synthesis with a fairness-aware pre-processing transformation. To our knowledge, this is the first work to attempt such an approach. The output of SAFES is a synthetic dataset with theoretical privacy guarantees that is adjusted via the reduction of structural bias to improve fairness for downstream classifiers. SAFES has several benefits. First, it is fully tunable with regards to both the privacy guarantees and fairness constraints. Second, for tight fairness constraints, SAFES exhibits fairness robustness measured by various metrics across a wide range of privacy guarantees per our empirical results, implying that one can adjust the balance between privacy and utility without a significant sacrifice in fairness. Third, though our examples and experiments focus on one commonly used DP data synthesizer and a well-known fairness-aware data transformation method, SAFES is a general framework that can admit different DP synthesizers of various DP guarantees and different fairness-aware data transformations satisfying various fairness metrics.

Translational Evidence: To illustrate the effective deployment of SAFES for solving real-world problems, we combine AIM, a graphical model-based DP data synthesizer, with a popular fairness-aware data pre-processing transformation. We run experiments on the Adult and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) datasets. The Adult dataset is a subset of US 1994 Census income data for 48,842 individuals that both contains privacy-sensitive variables (e.g., income) and encodes discrimination (e.g., pay disparities based on gender/race). The COMPAS dataset contains the criminal history information for 6,172 defendants and was used to predict whether an individual recidivates. Inferring membership in this dataset is a privacy violation as this is equivalent to determining that an individual was accused of a crime. Fairness concerns, such as racial bias in recidivism prediction, are also present. We evaluate the three-way privacy-fairness-utility trade-off for the SAFES procedure on each of these datasets. Privacy is measured via the privacy budget parameters used to generate the DP synthetic data. To evaluate fairness, we measure the change in structural bias for the synthetic data, as well as several commonly used fairness metrics for downstream classification tasks trained on the synthetic data (e.g., statistical parity difference, average odds difference, etc.). The general utility of the synthetic datasets is measured by comparing them to the original via the total variation distance (TVD) and the Kolmogorov-Smirnov (KS) test for distributional similarity. We also examine the ML utility the data by comparing several standard classification metrics on a logistic regression classifier trained on training data (with SAFES applied) and evaluated on test data (without SAFES applied). Substantial empirical evaluations for both the Adult and COMPAS experiments demonstrate that for reasonable privacy loss, SAFES-generated synthetic data achieve significantly improved fairness metrics with relatively low utility loss, implying that SAFES can be effectively employed to mitigate privacy and fairness concerns simultaneously for fully synthetic data while preserving good data utility for downstream learning tasks.

RISE Dimensions: The SAFES procedure aligns nicely with the RISE principles, especially the "responsibility" and "ethics" dimensions. Privacy and fairness are well known ethical concerns that surround the current AI climate. When building and deploying AI models, it is (or should be) the responsibility of those building the AI models to ensure that such widespread concerns are properly addressed. In this context, the SAFES procedure represents a novel and useful tool for those who seek to be more responsible in their AI practices in these areas; it can be an effective way to simultaneously address privacy and fairness concerns for real-world data and has the potential to positively impact society in a wide range of fields (e.g., healthcare, hiring, criminal justice).

#60: Explanation Difference: An Equitable Approach Towards Fair Machine Learning

Authors: Joe Germino, Yuying Zhao, Tyler Derr, Nuno Moniz and Nitesh V. Chawla

Problem Statement: Thus far, Fairness in Machine Learning (ML) has been focused entirely on the equality of predicted values across protected groups without regard for how a model arrives at these predictions. This can lead to a situation where a model which is attempting to correct for societal biases through its predictions, instead creates different incentive structures for members of different protected groups further perpetuating these biases. This traditional approach is focused on measuring the equality of predictions without concern for the equity of the model. We illustrate the urgency in considering a model's decision-making process when measuring fairness. We demonstrate the inequity of traditional fairness approaches and argue that fairness should be treated as a multi-objective optimization problem in which both prediction and explanation disparity are treated equally.

AI Innovation: We propose a new fairness measure, Explanation Difference (ED), and demonstrate how it can be efficiently integrated within existing optimization techniques. Specifically, we utilize FastSHAP to measure the disparity in explanations between counterfactual samples focusing on the disparate treatment of two similar people. We argue that fairness should be reframed as a multi-objective optimization task in which predictive performance, traditional fairness, and explanation fairness all receive weight.

Translational Evidence: Our results show that, used in tandem with existing fairness measures, ED can lead to more equitable fairness. We demonstrate that existing fairness-aware algorithms are capable of optimizing for equality of predictions but they do so with biased explanations. With our approach, we are able to optimize a model to be competitive in both predictive performance and traditional fairness with existing fair algorithms while improving the explanation disparity.

RISE Dimensions: Our proposed framework has broad implications for the deployment of Responsible AI. Fairness is a key component in real-world scenarios to prevent harmful stereotypes and societal biases from influencing our decision-making. Our approach builds on traditional, equality-based fairness and explores a new equity-based concept. We believe that using our multi-objective framework will lead to overall fairer ML systems.

#74: Reinventing the Global Computing-Energy-Land Nexus to Integrate Artificial Intelligence into Climate Action

Authors: Pengyu Zhong, Wenze Yue, Zhehao Liang, Sisi Meng and Tianyu Wang

Background: AI is Eating the World's Energy and Land Artificial Intelligence (AI) is reshaping digital earth while unevenly underscoring the impact of climate change. Substantial computational power demand triggers an unprecedented increase in energy consumption and associated greenhouse gas emissions. According to the International Energy Agency (IEA), data centers currently account for approximately 0.6% of global carbon emissions and consume between 1% to 1.5% of worldwide electricity, which are projected to double within the next two years, driven by AI and cryptocurrencies. To address the enormous climate change, Sustainable Development Goals (SDGs) emphasizing renewable energy. Among these, solar photovoltaic (PV) technology emerges as a critical component of climate action. Currently, solar PV constitutes approximately 3.6% of global energy production, becoming the third-largest renewable energy source globally. This rapid expansion of solar PV infrastructure, however, has introduced an emerging environmental conflict. In particular, croplands are well suited for PV installation due to morphological and climatic reasons, for instance, half of the areas in the western U.S. that are suitable for PV installation are also suitable for agriculture. Therefore, a key challenge to building resilient nexus Computing-Energy-Land (CEL, Fig.1). This article proposes a holistic approach to reimagining the CEL nexus. First, we identify global internet data centers (IDC) through an AI image recognition based on YOLO-3. Second, a model for solar panel suitability that incorporates the panel's microclimate derived from first principles and social-economic indicators. Third, we further utilize counterfactual framework and estimate the climate mitigation potential of reallocating IDC integrating land carbon sequestration and in contrast with the baseline scenario and different RCPs situation.

Methods: First attempt to identify reallocating IDC potential for climate change This article is the first attempt to address the gap of limited information on climate change impacts of incremental resource consumption due to IDC expansion for AI. The working flowchart is shown in Fig.2. There are three main steps: First, two-stage processes is performed to identify IDC. To address the complex characteristics of the spatial distribution of IDCs, (1) a top-down approach using YOLO-3 learning to identify the area of IDC through visual interpretation using optical remote sensing image. (2) bottom-top stage refers to extrapolate IDC area globally combining more social-economic data to consider the economic revenue of IDC. Second, after requiring effective area of IDC, we focus on PV as the main alternative for the energy transition, which is the most competitive with land use and other renewable paths like wind power can be located in far offshore areas to avoid land consumption. In this step, we generate global PV installation suitability derived from heat balance principle for microscopic photovoltaic panels.

Current result and discussion: Embedding AI in broader SDGs and combined policy implication We have obtained the typical data center characteristics around the world, and next we need to use deep learning and extrapolate the prediction of the global IDC distribution area to reduce the uncertainty in IDC recognition. Meanwhile, we have used global climate data for exploratory analysis to generate a global PV efficiency distribution map. A potential inverted U-shaped relationship between PV conversion efficiency and land carbon sequestration was observed using theoretical emission reduction abatement potential. This suggests that the relationship between PV installation and land carbon sequestration follows a law of diminishing margins. Research is critical for developing combinatorial policies to achieve inclusive governance in AI era.

#89: DesignPilot: Mitigating AI Hallucinations in Automated CAD Model Generation through GCN-Transformer Guidance

Authors: Shifu Hou and Fanny Ye

Large Language Models (LLMs) have shown significant potential across various domains, but when deployed as AI copilots in complex, specialized tasks, they often generate outputs that appear plausible yet are inaccurate due to AI hallucinations. To overcome this challenge, we propose DesignPilot—an intelligent system that seamlessly integrates LLMs with computer-aided design (CAD) to automatically generate CAD models. Our approach employs a novel step-by-step planning method combined with a Graph Convolutional Network (GCN) and Transformer-based link prediction retrieval model to enhance the accuracy and reliability of generated designs. The core workflow of DesignPilot includes step-by-step planning, GCN-Transformer link prediction retrieval, LLM-driven code generation, and model validation using an advanced CAD simulation engine. In this workflow, the GCN-Transformer model analyzes the relationships between different steps in the CAD design process to predict and retrieve relevant example code, effectively guiding the LLM in generating CAD model code that adheres to physical constraints and design specifications. Experimental results demonstrate that DesignPilot significantly reduces the risk of AI hallucinations by closely aligning LLM outputs with actual design constraints, thereby improving the success rate and accuracy of CAD model generation and paving the way for future advancements in AI-assisted engineering design.

#106: Benchmarking LLMs on Improving Coding Efficiency

Authors: Vince Andriacco, Zach Petko, Mengzhao Jiang and Meng Jiang

Large Language Models (LLMs) are increasingly used in software development to generate, translate, and optimize code. As their role grows, robust benchmarks are essential to evaluate their real-world effectiveness. Most existing benchmarks focus on code correctness rather than efficiency. The few that do consider performance often target only Python or use artificial tasks like LeetCode, which are far removed from real-world programming scenarios. However, optimizing code efficiency across multiple languages is crucial for practical software development. This highlights the need for a benchmark that evaluates LLMs in realistic, performance-critical scenarios. In this work, we propose a holistic benchmarking framework to assess LLMs in both intra-language optimization and cross-language translation with a focus on performance. Our benchmark covers five widely used languages—C, C#, Python, Java, and R—and is built from actively maintained, well-documented open-source GitHub repositories to ensure relevance to real-world coding

applications. We evaluate models using consistent metrics such as execution time, memory usage, and code size, tested in Docker-based environments for reproducibility. This work aims to establish a practical and scalable benchmark for assessing the performance optimization capabilities of LLMs in real-world coding scenarios.

#110: Synthetic Data with Heterogeneous Differential Privacy

Authors: Regina Mannino and Fang Liu

Problem Statement: As AI and technology continue to expand and evolve, the challenge of protecting individuals' data becomes even more urgent. Differential privacy (DP), a framework that mathematically ensures privacy for individuals, offers one solution to this complex problem. DP typically provides a uniform privacy loss parameter to an entire dataset which ensures that every individual and variable is protected with the same privacy guarantees. While a uniform privacy loss parameter can be considered the fairest option, individuals often have varying privacy preferences. Furthermore, certain variables may be highly sensitive and need stronger privacy guarantees than less sensitive ones. Overall, the uniform DP framework might not be optimal for the tradeoff between privacy and utility. Allowing for user- and attribute-specific privacy loss parameters can address the issue of limited utility due to DP guarantees, while also allowing users more input on how their data is used and how they want information to be protected. Previous literature has focused on applying one of the two dimensions (user, attribute) at a time; we propose a new framework that combines both cases to allow for more flexibility while also improving utility.

AI Innovation: We introduce multi-dimensional heterogeneous DP (MUD-HDP) by combining user-dependent DP (UD-HDP) and attribute-dependent DP (AD-HDP) guarantees according to a privacy budget allocation policy. We propose a privacy allocation policy to combine competing privacy requirements of UD-HDP and AD-HDP, a main challenge of merging these two frameworks. This can be extended to or adjusted for other scenarios based on domain knowledge and user requirements. We also address how diverse privacy needs can harm the utility of the results, and we develop a softmax weighting scheme to address this issue. This allows us to down weight the contribution of highly perturbed privacy groups at small sample sizes by borrowing information from less perturbed groups to improve the overall utility. A popular approach for releasing information under privacy guarantees is to release individual-level synthetic data. Since DP guarantees can be enforced during the creation of the synthetic dataset, repeated analysis can be performed on the synthetic data without incurring additional privacy loss. To our knowledge, our work is the first to incorporate HDP in synthetic data generation. We propose a synthetic data generation method for user groups with diverse privacy needs and across attributes with different levels of sensitivity and release multiple synthetic datasets, allowing for proper uncertainty quantification and valid statistical inference about the model parameters of interest. Overall, this framework facilitates responsible and effective AI research and discoveries by allowing analysts to train models on rich synthetic data while upholding rigorous privacy standards. Advances in DP and synthetic data are especially valuable for researchers in other fields seeking to share data, with expanded access to high-quality synthetic datasets enabling more privacy-conscious practices across disciplines.

Translational Evidence: To compare UD-HDP, AD-HDP, MUD-HDP and uniform DP on the utility of the synthetic data generated in each framework, we conduct extensive simulation studies for various data types and a real data set on obesity. We split the sample into three disjoint privacy groups (conservative, moderate, and liberal) with varying privacy parameters and consider a mixture of private and public attributes. We assess the performance of each method in parameter estimation and inferences under DP constraints. In general, we see improvement in the inferential results when softmax weighting is employed. For the univariate case, UD-HDP outperforms uniform DP. As the number of variables increase and we can incorporate AD-HDP and MUD-HDP, where MUD-HDP performs better or similarly to AD-HDP and outperforms UD-HDP and uniform DP in nearly all data scenarios. Overall, incorporating HDP improves the utility of results from DP synthetic data generation.

RISE Dimensions: Data privacy is fundamental to the discussions surrounding responsibility, integrity, and ethics in AI. AI models are often trained on datasets containing sensitive information, requiring researchers and companies to responsibly handle data privacy through robust frameworks such as DP or by sharing synthetic data. HDP gives users an avenue for specifying how they want their data protected, and companies or researchers in charge of the data uphold integrity by honoring user preferences in the data generation process, leading to released data and information of better

utility while respecting user input. Furthermore, HDP, compared to the one-size-fits-all uniform DP approach, offers a fairer and more nuanced approach, allowing more granular privacy controls. This not only better serves both the needs of individuals and analysts but also encourages critical reflection on who defines which data should remain private.

#114: Robust Machine Unlearning

Authors: Dongwhi Kim, Katherine O'Roark, Zheyuan Liu and Meng Jiang

As machine learning expands into critical domains like healthcare, finance, and personal devices, the privacy of training data has become increasingly important. Machine unlearning has emerged as a vital technique to address these concerns, particularly in light of regulatory frameworks such as the EU's General Data Protection Regulation (GDPR) and the California Privacy Rights Act (CPRA). These regulations establish a "right to be forgotten", mandating mechanisms for individuals to request the removal of their personal data from systems, including AI models trained on such information. Despite this growing legal and ethical necessity, the current landscape of unlearning algorithms fails under adversarial attacks, which can recover up to 88% of the supposedly "forgotten" information. We propose to develop novel unlearning techniques that maintain effectiveness against adversarial attacks while preserving model utility and generalizability. Our approach tests on two benchmarks: MUSE and WMDP to evaluate and improve unlearning robustness.

#134: Interpretable Latent Space Disentanglement in Generative Models via Feature Variance Heatmap with Latent Traversal (FVH-LT) and Dirty Block Sparse Regression (DBSR)

Authors: Xiaoan Lang and Fang Liu

Problem Statement: Understanding the internal and interpretable latent representations learned by end-to-end deep learning models remains a fundamental challenge in trustworthy machine learning and AI due to their black-box nature. This lack of transparency hinders the capability of practitioners to validate or interpret model behavior and to make the best use of what the models have to offer, especially in critical applications where model-based decisions must be explainable and accountable. Disentangled representation learning offers a promising solution by attempting to learn independent and semantically meaningful generative factors from data. Leveraging the framework of variational autoencoders (VAEs) is a popular approach to learning such disentangled representations. However, despite significant progress, most VAE-based disentanglement methods face several limitations. First, they often rely on access to ground-truth generative factors for evaluation—information that is rarely available in practice, Second, they are predominantly developed on image datasets, with limited exploration on their applicability to and other data types, such as tabular data. Third, these models commonly suffer from label switching, where the indexing of latent dimensions varies across training runs, hindering reproducibility and interpretability. To address these challenges, we extend existing VAE disentanglement frameworks by introducing two novel techniques -- Feature Variance Heatmap with Latent Traversal (FVH-LT) and Dirty Block Sparse Regression (DBSR) -- for interpretable latent space analysis that do not require supervision or prior knowledge of generative factors. Furthermore, we propose an alignment strategy based on greedy matching to address the label-switching problems, enabling consistent interpretation of latent dimensions across multiple training runs. Finally, our techniques apply to diverse data types and modalities.

AI Innovation: To move beyond the limitations of existing methods, we develop two techniques -- FVH-LT and DBSR -- for interpretable latent space analysis. Both techniques are fully unsupervised and applicable across diverse data types—including tabular and image data—making them broadly accessible in practice. FVH-LT improves interpretability by performing latent traversal on each latent dimension in the embedding space learned through an encoder given the input features —varying one at a time while keeping others fixed — and generating reconstructions through a decoder. By analyzing how the reconstructed features vary in response to changes in individual latent dimensions, FVH-LT captures the relations between each latent variable and the input features and its influence on the generated data. This process results in a variance heatmap that helps to interpret the learned latent space. The second technique we propose is DBSR, which achieves the interpretability goal by identifying the association between input features and latent dimensions through a multi-task sparse regression framework. After training the VAE, the posterior means of the latent

variables are extracted and treated as response matrices, while the original input features serve as design matrices in a multi-output regression model. DBSR decomposes the coefficient matrix into a block-sparse component that captures shared structure across latent dimensions and an element-wise sparse component that identifies feature-specific associations. The element-wise sparsity provides interpretable associations between input features and their latent variables while simultaneously serving a feature selection role by identifying which the most relevant features to each latent dimension. In both FVH-LT and DBSR, the latent dimensions with higher KL divergence between their prior and posterior distributions tend to encode more disentangled and meaningful representations. To distinguish these from noisy or uninformative latent dimensions, we apply a clustering-based strategy to the KL divergence for the latent dimensions. To optimize the stability in the dismantlement results, we suggest running the method (either FVH-LT or DBSR) multiple times and aggregating the results from the multiple runs. This would lead to the label-switching phenomenon where latent dimensions are inconsistently indexed across the multiple runs. To address this issue, we apply a greedy alignment procedure based on correlations among informative latent dimensions. Collectively, all the above technicality innovations form a general-purpose, unsupervised framework for interpreting VAE latent spaces that is applicable to a wide range of data types. To our knowledge, this is the first work on interpretable latent space disentanglement in VAEs that would be not require users to possess especially within the underexplored context of tabular data. our work aims to advance the development of interpretable representation learning methods that generalizes across, especially in data-constrained environments.

Translational Evidence: To assess the performance of FVH-LT and DBSR on learning disentangled interpretable latent representations, we conduct extensive experiments in various settings, including a benchmark image dataset (the MNIST dataset), simulated tabular data (generated from a factor analysis model with pre-defined latent factors), and two real-world benchmark tabular datasets (the white wine quality dataset and the FIFA 2018 World Cup match statistics dataset). In the MNIST experiment, FVH-LT identifies a subset of latent dimensions with high KL divergence that exhibit significantly higher pixel-wise variance in the heatmaps. These informative latent dimensions reveal clear digit-like spatial structures, highlighting potential disentangled representations without using the label information or the ground-truth knowledge of the generative factors in the MNIST dataset. For the two simulated tabular datasets, both FVH-LT and DBSR successfully recover the underlying generative factors, with individual informative latent dimension aligning closely with distinct generative factor blocks. These results validate that both methods can capture semantically meaningful patterns without labels. For the two real-world datasets where the true generative factors are unknown, both FVH-LT and DBSR consistently recover latent dimensions that correspond to interpretable feature groups. These features reflect interpretable aspects of wine quality and soccer match performance, respectively, and demonstrate each method's ability to learn meaningful structure in complex tabular data. Collectively, these findings suggest that FVH-LT and DBSR demonstrate real-world promise and viability for learning interpretable disentangled latent space across different data types and practical domains.

RISE Dimensions: Limited interpretability of deep generative models such as VAEs remains a major obstacle to building trustworthy AI systems. Our proposed techniques — FVH-LT and DBSR — directly address this challenge by making the latent space of VAEs more transparent and understandable. By revealing how latent dimensions relate to input features, these methods empower practitioners to inspect and validate what the model has learned, promoting responsibility and fostering trust in generative AI deployment. Notably, both FVH-LT and DBSR are unsupervised, require no ground truth knowledge on generative factors to validate the finding, and are applicable to different types of data. The robustness and flexibility of these techniques align with the general principle of representation learning, helping to enable responsible AI analysis in under-explored domains without rich labeled training data.

#135: Decoding the Carbon Cost of AI: Toward Transparent and Equitable CO_2 Emissions Estimation for LLM Inference

Authors: Vedanth Nandivada, Alexi Orchard and John Behrens

As the adoption of Large Language Models (LLMs) accelerates across education, healthcare, business, and government sectors, concerns about their hidden carbon footprint have emerged as a major global sustainability challenge (Luccioni et al., 2024; Ren et al., 2024). Although LLMs are celebrated for their capabilities, the emissions associated with their use—particularly during inference, when models generate outputs—remain largely invisible to end-users, policymakers, and institutions. Current tools for estimating these emissions, such as MLCO2 (Lacoste et al., 2019), simplify environmental costs by relying primarily on infrastructure variables like hardware type, data center location, cloud provider, and time spent prompting. However, these methods often fail to capture the full complexity of usage patterns and system-level behavior. Newer frameworks like LLMCarbon (Faiz et al., 2024) offer deeper modeling of emissions tied to hardware architecture and computational strategies, yet real-world benchmarking remains limited. In this project, we address the urgent need for more accurate, transparent, and actionable emissions estimation for LLM deployment.

To explore the drivers of carbon emissions from LLM inference, we built predictive models grounded in emissions data generated using MLCO2, complemented with detailed usage behavior metrics. The dataset captured system-level factors (hardware chip type, data center region, provider, time spent prompting) and usage-level factors (number of tokens generated, number of prompts, average response length, response time per prompt, number of sittings, and cumulative prompting time until task failure, or "Find the Limit" [FTL]). We conducted a baseline linear regression analysis and two random forest regression analyses to identify key predictors of emissions.

Early results showed that system-level factors, especially hardware and time spent prompting, dominate emissions behavior, while usage patterns provide important secondary nuance. In future work, we plan to benchmark emissions estimates using LLMCarbon's methodology, comparing results to MLCO2-based predictions to highlight strengths and limitations across approaches. This two-method benchmarking will support the development of a more explainable, auditable emissions framework for LLM deployments. All emissions data were generated through structured simulation of LLM prompt sessions, reflecting a range of prompting times, output lengths, and interaction patterns. Random forest models were trained to identify nonlinear dependencies between usage behavior and emissions, while linear regression served as a baseline comparison. Dataset provenance and modeling processes will be detailed in future open-access materials to enhance reproducibility and support broader scientific engagement.

This research lays the groundwork for building practical emissions auditing tools for universities, enterprises, and public institutions deploying AI. For example, a university IT department could simulate procurement scenarios using the estimator to prioritize sustainable hardware and cloud vendor choices. Data center operators could use refined emissions projections to design smarter, lower-carbon infrastructure. Policymakers could integrate these tools into AI-specific environmental reporting standards to ensure more responsible technology growth. In future stages, we plan to share the developed emissions estimation framework, enabling broader public use, external validation, and cross-sectoral feedback.

This work advances Responsibility by exposing hidden environmental costs of AI systems; Safety by highlighting long-term ecological risks from large-scale deployments; Ethics by calling for transparency in infrastructure decisions that affect public resources; and Inclusion by empowering historically marginalized communities to challenge datacenter siting decisions with better emissions evidence. By building tools that make environmental impacts visible and actionable, this project supports environmental justice and promotes more equitable AI development across global contexts. By integrating methods from machine learning systems analysis, environmental science, and public policy evaluation, this project demonstrates how technical rigor and societal responsibility can be jointly advanced. The approaches developed here could form a model for carbon transparency standards not only in LLMs, but across other AI domains such as computer vision, robotics, and speech recognition.

References:

- Faiz, Ahmad, et al. “LLMCarbon: End-to-End Carbon Footprint Projection of LLM Inference." arXiv preprintarXiv:2402.01205 (2024).
- Lacoste, Alexandre, et al. “Quantifying the Carbon Emissions of Machine Learning." MLCO2.org (2019).
- Luccioni, Alexandra Sasha, et al. “Power Hungry Processing: Watts Driving the Cost of AI Deployment?” ACM FAccT 2024 (2024).
- Ren, Shaolei, et al. “Reconciling the Contrasting Narratives on the Environmental Impact of Large Language Models." Scientific Reports 14 (2024): 26310.

#149: Responsible Deployment of AI-enabled Drones in Policy and Practice

Authors: Demetrius Hernandez and Jane Cleland-Huang

Problem Statement: Uncrewed aerial vehicles (UAVs) are evolving from remotely piloted tools into highly autonomous, AI-enabled platforms that can save lives in disaster zones, inspect aging bridges, and monitor fragile ecosystems. Exactly the same advances in onboard perception, navigation, and teaming, however, can be (and already are) re-purposed for mass surveillance, political repression, and lethal autonomous strikes. This dual-use dilemma is compounded by human-machine teaming: as autonomy increases, accountability and operator oversight become harder to guarantee. Current governance frameworks dwell on pre-deployment regulation and overlook the moment-to-moment decisions that human operators must make while missions unfold. The grand challenge, therefore, is to devise technical and policy mechanisms that (1) unlock the civilian benefits of AI-enabled drones, (2) curtail malicious or reckless uses, and (3) embed Responsibility, Inclusion, Safety, and Ethics (RISE) across the technology’s full life-cycle.

AI Innovation: We merge two complementary lines of work to meet this challenge. - First, analysis of AI-enabled drones maps the technology stack from edge computing and swarm coordination to lethal autonomous weapons and distills the risk landscape across autonomy, algorithmic bias, automation bias, and proliferation pathways. This analysis yields a system-level view showing where legal, technical, and organizational controls must interlock. - Second, we introduce RAVEN (Runtime Advocate Views for Event-driveN personas), a novel runtime-requirements framework that extends static user personas into adaptive, event-driven advocates. Three advocates: a Safety Controller, an Ethical Governor, and a Regulatory Auditor, continuously translate evolving world states into just-in-time guidance for the human operator. Implemented with an in-context, standards-grounded large language model pipeline, RAVEN surfaces safety, ethical, and regulatory information to the operator exactly when they matter, closing the gap between policy intent and operational reality.

Translational Evidence: Autonomous drone systems offer a highly adaptable platform to support critical societal needs such as public health, food security, clean energy, resilient infrastructure, and climate adaptation. The same drone swarm that delivers blood supplies to rural clinics can be reconfigured to monitor crop conditions, inspect solar installations, or map high-risk flood zones, making it a uniquely scalable tool for advancing well-being, sustainability, and urban resilience. By embedding RAVEN’s Safety, Ethical, and Regulatory advocate personas directly into each mission is guided by real-time operational information that upholds principles of privacy, accountability, and public safety. It enables responsible drone use across sectors (healthcare, disaster response, agriculture, and infrastructure) while lowering regulatory friction and minimizing unintended harm. In doing so, this work helps transform broad development goals into actionable, auditable decisions at the point of use.

RISE Dimensions: Responsibility is woven through analysis of dual-use risk and RAVEN’s Regulatory Auditor, which anchors every autonomous action to traceable standards such as FAA Part 107 and MIL-STD-882E, closing accountability gaps in human-machine teams. Inclusion is promoted by use cases that prioritize remote medical delivery, disaster relief, and environmental monitoring, directing AI-driven benefits toward underserved and high-risk communities rather than exclusively militarized or urban markets. Safety is enforced by the Safety Controller advocate, which continually checks environmental, system, and collision margins. Ethics is embedded via the Ethical Governor, filtering privacy-intrusive sensor use, flagging bias, and highlighting unethical functions, thereby translating abstract principles into real-time operational guardrails.

#152: Life or Death Chemistry: When 1 in 6 Chemotherapy Medications Fail Quality Tests

Authors: Christopher Sweet, Priscila Saboia, Maximilian Wilfinger and Marya Lieberman

In sub-Saharan Africa, cancer patients face a silent threat: substandard chemotherapy medications. Our NIH-funded study revealed that 16.8% of tested chemotherapy drug batches failed to meet the United States Pharmacopeia quality standards. While detection of completely falsified drugs has advanced, a more critical challenge remains: determining whether medications contain the proper concentration of active ingredients. Diluted chemotherapy drugs appear legitimate in basic tests but deliver ineffective treatment doses. Traditional concentration testing methods remain inaccessible in low-resource settings, creating a critical need for point-of-care technologies. This paper presents our AI-driven solution that quantifies chemotherapy drug concentrations at the point of care.

Bridging Detection and Concentration: An Interdisciplinary AI Approach

Our international collaboration—spanning institutions in the United States, Malawi, Ethiopia, Cameroon, and Kenya—brings together expertise in pharmaceutical chemistry, analytical science, computer vision, and machine learning. Building upon the Paper Analytical Device (PAD) platform—a low-cost, chromatography-based card with 12 reagent lanes that detects the presence of drugs—we have created an AI-driven system that additionally quantifies drug concentration from subtle variations in color patterns. This computational pipeline transforms qualitative visual assessment into precise quantitative analysis by extracting RGB values from PAD images, applying Partial Least Squares (PLS) regression to identify concentration-relevant features, and employing neural networks to capture non-linear relationships in color responses.

The system integrates seamlessly with a smartphone application designed for use in remote clinics, providing healthcare workers with immediate, interpretable concentration results alongside confidence metrics. This "bench-to-bedside" model demonstrates how sophisticated AI can become accessible in environments previously beyond the reach of advanced computational tools, connecting laboratory science directly to patient care in regions where traditional testing infrastructure is unavailable.

A Hybrid Mathematical Framework for Resource-Limited Settings

The mathematical foundation of our approach addresses the extraction of concentration information from high-dimensional, noisy image data. Each PAD card generates color patterns, resulting in hundreds of features (RGB values). Our innovation uses a two-stage hybrid architecture: First, PLS regression projects the high-dimensional color data onto a lower-dimensional latent space: t = Xw (projection of X into latent space), u = Y c (projection of Y into latent space). This reduces the feature space to 5–20 latent variables capturing relevant color variations. Second, these latent variables feed into a neural network: Concentration = NN(t1, t2, ..., tk). This approach addresses the "curse of dimensionality" by learning in a compact, information-rich feature space rather than raw color space. It overcomes the critical limitation of pure neural networks—their need for extensive training data—by using PLS to create a chemically meaningful feature space that requires fewer examples to achieve accurate concentration predictions.

Our technical design prioritizes accessibility: low cost (under $2 per test versus $50+ for conventional methods), minimal training requirements (guided by the smartphone interface), rapid results (under 5 minutes versus days for laboratory testing), and offline operation without specialized infrastructure. This transformation moves sophisticated pharmaceutical analysis from centralized laboratories directly to point-of-care settings where treatment decisions are made.

From Theory to Practice: Validation and Future Vision

We validated our approach using over 4,000 PAD images of six chemotherapy drugs at three concentration levels (100%, 66%, and 33%). Our neural network achieved 93.4% drug identification accuracy—reaching over 99% when accounting for chemical similarities between platinum-based compounds. For concentration determination, our hybrid approach, applied to our previous FHI360 data set, reduced the mean square error by approximately 40% compared to PLS regression alone (from 15% to 9% MSE), demonstrating the value of capturing non-linear relationships in colorimetric data. Validation revealed that, for the ChemoPAD data set, lighting conditions during image capture significantly impact the accuracy of concentration prediction—a critical insight for field deployment. These discoveries

inform our next-generation PAD design with integrated color reference standards for automatic lighting compensation. Our vision extends beyond chemotherapy to antimalarials and antibiotics, creating a sustainable cycle of field data collection, AI refinement, and device evolution. By involving local institutions in data collection and validation, we build capacity for independent innovation in regions traditionally excluded from cutting-edge AI development. This work exemplifies RISE principles: Responsibility, through rigorous validation to ensure trustworthy clinical use; Inclusion, by co-developing solutions with underserved communities; Safety, through broad-spectrum drug and dose testing; and Ethics, by addressing systemic healthcare inequities. Our iterative, collaborative model creates a scalable blueprint for equitable, AI-driven pharmaceutical quality control.

#162: Empirical Bayes Tensor Decomposition: A Holistic and Interpretable Representation of Digital Trace Patterns

Authors: Xinyuan Zhang, Junhui Cai, Jingjing Li and Ahmed Abbasi

The increasing availability of digital trace data provides both great opportunities and challenges for modeling, understanding, and predicting user behavioral patterns and relationships with downstream outcomes. By trace data, we refer to time-stamped event log data, usually at the individual level. Some common examples of digital trace data include electronic health records, social media activities, and consumer clickstream records. Patterns derived from such data can help predict future near-term behaviors, shed light on user motivations and intent in digital environments, and detect mental disease. A fundamental challenge for studying user behavior through such data is disentangling digital traces from digital exhaust while maintaining representational richness. Digital trace data is longitudinal data containing some signal related to a downstream dependent variable, shrouded in an abundance of noise. Moreover, the multi-dimensional complexity of digital trace---encompassing temporal data, omni-channels, and activity patterns---presents significant challenges in extracting meaningful behavioral patterns from a high-dimensional large-scale dataset.

When deriving patterns from digital trace data, there are traditionally three key considerations in capturing potentially meaningful individual behavioral characteristics: the channels through which individuals interact and engage in activities of different kinds, the granularity to quantify user behavior through those channels, and the resulting sparsity level in both the input trace data and the underlying behavioral patterns. Despite merits and operational utility of digital trace data utilization, current digital trace studies explaining and predicting user behavior are limited on these dimensions. First, many prior work has focused on single-channel session-level data. For instance, when modeling online customer journeys, previous studies often concentrate only on patterns from a single online retailer website. Similarly, most research on mental health diagnosis and prediction focuses on a single modality such as text or audio. The single-channel perspective oversees both explanation and predictive power of cross-channel behavior, found in omni-channel studies, though these are rare and have received limited attention.

Second, the longitudinal nature of digital trace data often allows high temporal granularity. Clickstream and social media data usually have timestamps at the second level, allowing highly granular interpretation of activities. However, existing measures for digital trace data are limited in capturing complex patterns. Many studies use aggregated activity counts over a specific window which represent trace data as a single scalar value, such as average time spent on social media daily. This aggregating paradigm fails to consider the rich “within-window" patterns. Learning representations based on multi-dimensional digital trace data afford opportunities to parsimoniously learn and induce behavioral patterns such as temporal trajectories, channel consumption clusters, and behavioral interactions across time and space.

Third, the level of sparsity in the underlying trace data has implications for what can be gleaned, how (e.g., types of operationalizations), and the efficacy and utility of derived variables (i.e., ability to extract signal from noise). In general, behavior patterns based on scalar aggregations are considered more conducive to sparse input trace data. However, as mentioned previously, aggregated measures oversee the richness in temporal patterns and interactions across dimensions. On the other hand, direct incorporation of highly sparse input into traditional methods (e.g., frequency-based approach like maximum likelihood estimates) can result in misleading conclusions.

These special features call for more careful and suitable processing methods with digital trace data. One intuitive way to represent users’ digital channel interactions over time is to use a multidimensional tensor. For instance, user i visited

website j at time t. Following this intuition, we focus on the essence of modeling user behavior through digital trace data and construct a three-dimensional user-channel-time tensor, encompassing user interactions with or under various channels that can be website pages or human-centric signal modalities, and quantified in terms of frequency counts, at a high-level of granularity (e.g., by hour or second). While this tensor representation retains information of high-level granularity, with higher data dimension, prediction and optimal decision-making become prohibitively expensive.%, both in terms of sample size and computation. At the same time, looking for patterns in a tensor is like looking for a needle in a haystack -- it is hard to interpret and explain customer behaviors. On a separate note, user digital trace tensors usually have several special features: (1) The entries are count data, such as numbers of visits to webpage during a given time period; (2) Many entries are zero, i.e., the tensor is sparse; (3) Digital trace data can often be rich but noisy, making it difficult to distinguish signals from noise. Hence, we ask the following question:

Are there interpretable models and efficient representation learning procedures for modeling digital trace that can summarize and understand granular user-channel digital interactions across time?

In this paper, we tackle this challenge by proposing a Bayesian tensor framework that effectively processes multi-dimensional, longitudinal, and sparse digital trace data and captures user behaviors via low-dimensional latent factors using a family of sparsity-induced priors. The latent factors from decomposition are interpretable, reflecting behavioral patterns in the literature as well as shedding lights on new patterns. Our framework encompasses, first, a tensor representation of the data of three dimensions: session, channel, and time. The framework takes the sparse nature of digital trace data into consideration by imposing sparsity with spike-and-slab priors. The key procedure of our framework is tensor decomposition. We introduce an empirical Bayes methodology to estimate the (sparse) usage intensity across channels and time as well as their interactions. Different from previous development on tensor decomposition methods, our method is specifically designed for handling sparse count data, for modeling user digital trace and downstream applications. The final steps of our framework include various reconstruction and utilization of decomposition outputs for possible downstream application purposes.

We summarize our main contributions:

1. We propose a new Bayesian tensor framework for digital trace data modeling, specifically designed for count data with high sparsity level. The crux of our framework is the low-dimensional latent factors that represent different types of user behavior traits and corresponding behaviors across channel and time. We induce sparsity by imposing the spike-and-slab priors on the latent factors. Our framework is flexible and allows for a large class of scenarios/priors.

2. Our framework allows for varying levels of sparsity exhibited in factors. This feature mimics the heterogeneous nature of behavior traits; as we show later, online activity patterns have drastically different intensities across time, rendering different sparse structures.

3. We propose an empirical Bayes tensor decomposition method (EBTD) by adopting the variational Bayes approach. We show that EBTD outperforms other tensor decomposition methods in latent factor estimation, tensor recovery, explanation, and prediction tasks under various scenarios.

4. We apply our framework to two case studies to demonstrate downstream application value. The first involves a large test bed spanning 4 million raw user clickstream actions across 21 months. We construct a user-channel-time tensor spanning 38,046 user-session journeys over 336-hour windows measured across 141 websites related to seven different types of online channels. We also apply our framework on a multi-modality dataset from remote mental health assessments with 144 patients. Each patient-session spans a 12-minute window, with speech-to-text transcriptions, audio, video-facial, and video-gaze information captured at one second time intervals across 50 modality measures, resulting in over 5 million trace data points in the patient-modality-time tensor. We decompose the constructed tensors using EBTD and demonstrate the interpretability and prediction power of our method in explaining and predicting behavioral outcomes.

5. Beyond explanatory and prediction power, we also conduct cost-benefit analysis for both scenarios to further demonstrate the potential business and social effect of our framework. In the customer journey case, we show

potential revenue increase from the business perspective. In the mental health assessment scenario, we demonstrate potential capital savings from the social resources perspective, i.e., hospital resource savings.

6. For academic implications, following the recent editorial on pathways for design research on AI, we offer two salient design insights. Firstly, We propose tensor representation and decomposition methods that address the interplay between the 3 key features of digital trace data: omnichannels, temporal granularity, and resulting sparsity. Our proposed method is designed specifically to best accommodate these considerations and thus, better captures latent behavioral patterns. Secondly, the Bayesian tensor framework has downstream application potential in prediction, description and/or explanation aspect, with managerial and societal implications.

#173: A Personalist Tech Ethics Framework for the Lithium-AI Frontier

Authors: Alejandro Williams Becker

I. Problem statement Global mining firms are beginning to test artificial-intelligence tools that predict lithium grades, fine-tune brine pumping, and cut water use. In Salta's Puna plateau—homeland of Kolla and Atacama peoples—public officials and community leaders confirm in recent dialogue sessions that they lack the vocabulary and conceptual tools to interrogate such algorithms once they arrive. The risk is not merely technical; it is moral. Without an ethical grammar that foregrounds persons-in-community, high-altitude deserts may again supply critical minerals while remaining epistemically peripheral to the very technologies they enable.

II. AI innovation Rather than offering another dashboard or optimization model, this paper proposes an ethical-governance framework that decision-makers can apply before industrial AI becomes routine. The framework braids three strands: A. Personalist-communitarian ethics recognizes every human person as a relational agent whose flourishing is tied to the common good; therefore, data, water, and algorithmic benefits must be distributed according to principles of participation, subsidiarity, and care for creation. B. Politics of artifacts (Langdon Winner) insists that AI systems embody power relations; making model assumptions explicit is a precondition for legitimate consent. C. Postphenomenology (Idhe, Verbeek) examines how AI mediates human-world relations and reshapes horizons of action; it invites situated reflection on how predictive models may re-configure Indigenous practices of land stewardship. Together these lenses yield a six-step deliberative protocol: mapping stakeholders, surfacing hidden technical choices, evaluating human-technology mediations, discerning community goods, negotiating shared metrics, and iterating in public.

III. Translational potential Applied to an upcoming provincial lithium project, the protocol would: A. provide a common language for engineers, policymakers and community assemblies to discuss AI tools without requiring advanced coding skills; B. generate participatory criteria (e.g., "water-first optimisation," "explainability in Kolla") that can be written into permits and CSR agreements; C. offer universities and NGOs a scaffold for capacity-building workshops that raise AI literacy while respecting local epistemologies.

IV. RISE alignment Responsibility: embeds algorithmic accountability in permit negotiations through shared ex-ante criteria. Inclusion: centers Indigenous voices and bilingual discussion; protocol scalable to other critical-mineral regions. Safety: prioritizes water-risk mediation and continuous, community-led review of AI impacts. Ethics: marries personalist dignity with contemporary philosophy of technology, producing a culturally resonant governance tool.

V. Conclusion AI may one day make lithium extraction cleaner; this framework ensures it also becomes more just. By fusing personalist ethics, the politics of artifacts, and postphenomenology, Puna communities and policymakers can co-author the normative boundaries of the algorithms that will soon shape their land, water, and future—fulfilling the Responsible, Inclusive, Safe, and Ethical mandate of the RISE AI Conference

Track: Education and Workforce of the Future

#61: Empowering Language Learners with AI: A Carnival Music Workshop for Portuguese Students

Authors: Isadora Teles de Oliveira Gouveia

With the rise of AI usage in multiple societal settings, especially at the educational level, it has become essential for educators and institutions to address the question of AI in the classroom. As a language instructor, tools such as chatGPT and others have become incredibly popular among students when doing their foreign language assignments. This poses the question of how to incorporate these tools in the language classroom effectively, in a way that contributes to their language improvement, promotes higher intercultural awareness, and expands their knowledge about cultural practices in the countries that speak the target language. To achieve that, it is interesting to look at the use of AI to generate content that is culturally relevant to students.

This project aims to promote student engagement and community building through the use of AI to generate a carnival song in Portuguese. The study was conducted in Massachusetts, home to many Portuguese-speaking immigrants and heritage speakers; therefore it is usual to have carnival celebrations in the region. Portuguese learners from the second semester had a lesson about how carnival is celebrated in Lusophone countries, such as Brazil, Portugal, and Cape Verde. Students were invited to put these cultures in dialogue, by sharing their experiences with the festivity, as well as learning about musical genres and the history behind the lyrics of carnival songs, such as 'marchinhas' and 'samba-enredos', and how they play a part in the world-famous parades.

Following the talk section of the lesson, students participated in an in-class Carnival Music Workshop, in which they practiced Portuguese language skills by coming up with lyrics to their own carnival song. At this moment, as an instructor, I assisted them with their questions and students used AI tools such as ChatGPT for brainstorming ideas, rhyming suggestions, and language correction checks. After they crafted lyrics, they fed those into the AI tool UDIO.com, where they could select the length and characteristics of their songs, such as the musical genre and the gender of the singer.

After they worked on their songs, the class heard a couple of samples of the songs generated. An in-class discussion was held then on the significance of AI usage to generate music, with the guiding questions: “Do you think this sort of AI usage could be incorporated by songwriters to write their samba-enredos?”, “What do you think the future of music looks like with these technologies on the rise?” and “In your opinion, are there any ethical implications of using AI tools to generate music? If so, what are they?” Students had mixed opinions about using AI to generate music, with some sharing their concerns about the lack of originality and “soul” to the music process, whereas others pointed to saving time and money as benefits of AI music generation, making the process more accessible to the general public. A questionnaire was conducted after the class to assess students' perceptions of the project regarding the cultural relevance, the usage of AI, and its implications.

Overall, we concluded that this hands-on activity allowed students to engage with an important aspect of the target cultures and, in the case of heritage learners, connect with their own cultural background. It also allowed for creativity in the exploration of different musical genres, as well as generating meaningful language practice and group work. Exercising critical thinking in the face of the rise of new technologies is also pointed out as a benefit of incorporating these activities in the language classroom. Lastly, this type of study is relevant to the field of language teachers because it can be easily replicated in any other target language and rhythm

#79: PipelineEDU: Creating High-Quality Synthetic Data for Educational Research and Applications

Authors: Austin Nicolas and Shahnewaz Sakib

Current educational AI research trains on large empirical datasets whose usage poses ethical concerns. Privacy loss risks leaking sensitive student information, including personal identifiers. To address the gap between educational AI research needs and ethically sourced data, this work generates a purely synthetic dataset with a replicable generation pipeline for usage in educational AI research. Purely synthetic datasets have no privacy loss since they are not drawn from empirical data. However the lack of similarity to real world data can impact the utility. The dataset generation pipeline creates a purely synthetic educational dataset containing student demographic information, educational history data, extracurriculars, career aspirations and future topics of interest. The dataset was built by sampling empirical distributions and weighted mappings between explicitly connected variables leading to significant correlations and distributions. However, gaps exist where empirical data to base the dataset on was limited. To explore the privacy utility tradeoff, the dataset was privatized using various methods based on differential privacy and shuffling. Classification models trained on the privatized data attempted to predict the student demographic information, representing the privacy loss. While, regression and regressification models predicted the career aspirations and future topics of interest, representing utility gains. Regressification is a novel technique that trains as a regression model but evaluates as a classification model. However, the dimensionality reduction technique limited the interpretability and utility of these results. Purely synthetic data supports educational AI research while minimizing ethical concerns around data sourcing, helping students and educators. Purely synthetic datasets allow educational machine learning models to be improved without compromising student privacy and safety

#118: Indiana's Comprehensive AI Support for K-12

Authors: Heather Herring and Matthew White

Poster moved to session on October 7th, 2025

Problem Statement: Indiana K-12 schools have varying levels of AI literacy and access to AI-powered tools at both student and teacher levels. Some schools have embraced AI initiatives, while others are now beginning that journey. The wide range of AI literacy and experience requires a broad spectrum of support and resources tailored to be responsive to the varied levels.

AI Innovation: In response to the identified problem statement, the Indiana Department of Education (IDOE) has taken a multipronged approach to increase AI literacy and access in K-12 schools across the state in conjunction with transparency around those efforts.

IDOE's Office of Digital Learning AI Pillars and corresponding areas of support include:

1. Messaging and Guidance

Guidance for schools - In April 2024, IDOE published key considerations related to AI and education for K-12 schools. Further expansion of the existing guidance will be released by Fall 2025. Parent guidance - Indiana's published AI Guidance for Parents and Families, composed of short videos, supports parents with key AI topics, including AI literacy and AI security.

2. AI Literacy AI Insights for Educators

A series of presentations from experts, researchers, and authors, featured important topics related to AI in education. Educators were able to attend live or watch recordings on demand. AI Spotlight Series - Educators across the state presented in the Indiana Learning Lab, the state professional development platform. Sessions from fifteen educators spotlighted practical ways they are using AI in the classroom. IDOE's AI literacy support is designed to support different groups, including beginners, advanced, technology leaders, administrators, and parents/caregivers.

3. Access

AI-Powered Platform Pilot Grant - In August 2023, IDOE invited accredited school corporations and school organizations to apply for a one-time competitive grant opportunity to fund a pilot of an artificial intelligence (AI) powered platform of their choosing for the 2023-2024 school year. This opportunity supported a one-year implementation and funded subscription fees and professional development to support high-dosage tutoring for students and reduce teacher workload through the use of an AI platform. This grant provided an opportunity for school corporations to target support to a specific building, grade level, subject area, or student population by integrating an AI platform with a cohort of teachers and students. Schools were encouraged to focus on student needs in response to academic impact data. The awardees consisted of 112 schools. This included 2,466 teachers and 45,244 students from 36 school corporations across the state. AI-Supported Alternative Education Grant.

In 2024, IDOE invited any approved alternative education program to apply for a one-time competitive grant opportunity to provide full-time enrolled students with access to an AI-powered platform intelligent tutoring system. This opportunity supported a one-year implementation of an AI-powered intelligent tutoring system that meets all vendor requirements for 5,333 students and 240 teachers at 24 alternative schools. Digital Learning Grant (DLG) - The annual Digital Learning Grant aims to support Indiana public school corporations and charter schools in their efforts to leverage technology to enhance learning experiences, foster innovation for students, and promote effective digital pedagogies for teachers. Artificial intelligence became a grant priority in the 2024-2025 school year and grants were awarded to support AI initiatives for 97,146 students and 8,941 teachers from 52 school corporations.

4. Data and Research Reporting AI Pilot research

In April 2024, IDOE launched a survey to collect feedback from all teachers who were part of the pilot grant and received 625 responses. The survey allowed teachers to provide direct input on their experience using the AI platforms. The primary goals of the survey were to understand platform usefulness, student and teacher impact, training effectiveness, and the overall experience using the platform. AI DLG research - To continue transparency efforts, IDOE is conducting another round of teacher feedback on their AI implementation. Findings from this feedback will be available late summer 2025.

Translational Evidence: By taking a comprehensive approach at the state level, IDOE is able to support - Return on investment - Our data and research reporting pillar provides transparency on time and funding investments through direct teacher feedback, which is ultimately used to make programmatic and policy decisions. AI-powered tutoring - For student-facing initiatives, the AI platforms are required to have certain capabilities, including a tutoring option for students. We recognize the immense value of a human tutor, but know there are fiscal and logistical limitations in scaling tutoring for students statewide. By providing access to AI tutoring platforms that meet privacy, safety, and support requirements, Indiana students have access to an AI tutor any time, and schools are able to target their own academic goals. Schools have shared positive anecdotes about the academic progress they are seeing. Large-scale AI literacy professional development - Through the Indiana Learning Lab, live and on-demand professional development sessions are able to reach 90k users, including teachers, administrators, parents, preservice educators, and more.

RISE Dimensions: The innovation aligns with the following dimensions: Responsibility and Safety - IDOE's efforts aim to prepare students for an AI-powered present and future - in a responsible way. The Office of Digital Learning's emphasis on increasing AI literacy for all is rooted in the drive to ensure AI use is responsible and safe. If education stakeholders at every level - from students to parents to administrators - understand what AI is, how it works, when to use AI, and what the risks are, those stakeholders are better equipped to support Indiana students

#144: EmpowerHER: A behavior change intervention to enhance breast cancer education and early detection among Latina women in Indiana, using LLM and ASR-enabled mHealth tools

Authors: Marcelo Guzman Aguirre, Dina Hanna, Judy Lee, Cheng Liu and Angelica García-Martínez

Latina women in the U.S. face a breast cancer paradox: despite lower incidence, they suffer higher mortality due to delayed diagnoses – exacerbated by poverty, limited English proficiency, and lack of preventive care. In South Bend, Indiana, where Latinos represent ~20% of the population, these disparities are intensified by structural barriers and cultural disconnects with the healthcare system. Our solution, EmpowerHER, integrates LLM-ASR technology within the SaludConecta mHealth system to address these challenges. The AI innovation combines large language models trained on region-specific Latina narratives with automatic speech recognition to deliver voice-enabled, culturally responsive breast cancer education. It also screens for mental health and SDoH using AI triage to tailor support and improve engagement.

Translational evidence from early pilots demonstrates improved comprehension, screening adherence, and appointment attendance among Latina users. Community co-design with La Casa de Amistad, and clinical delivery via Beacon and Saint Joseph Health mobile units, ensures scalable development. Aligned with the RISE framework, EmpowerHER prioritizes responsibility through algorithm audits, inclusion through bilingual, low-literacy design, safety via privacy-preserving architecture, and ethics through culturally grounded consent protocols. This model offers scalable, community-centered AI for addressing cancer disparities in underserved Latino populations.

#146: From Aspirational to Inspirational - Hoosier AI Maturity Model

Authors: Nida Ansari and Sundaresh Ramanathan

Problem Statement: According to McKinsey, artificial intelligence (AI) could add between $2.6 and $4.4 trillion in value annually. Deloitte states that 94% of executives believe that artificial intelligence will transform their industries in the next five years. Yet, research by RAND Corporation shows more than 80% of AI projects fail, wasting billions of dollars in capital and resources. There is an arc between awareness and implementation that most organizations fall on. We believe all organizations are aware, but don't know how to strategize, pilot, adopt, implement, and sustain. This demonstrates a need for a structured, systematic framework to help organizations to get to adoption and implementation with clear (real or perceived) risk and failure mitigation. The primary objective of this framework is to foster effective collaboration between Human and AI to maximize benefits across diverse sectors.

Real and Perceived Risks of AI Implementation

While there is general awareness and consensus on the need for ethics, fairness, and accountability with AI innovation, there is no clarity or consensus on the path forward. Organizations are faced with a number of unknowns:

- How will the current workforce be educated?
- What are the talent and workforce risks and costs?
- What is the cost to pilot and implement? How disruptive is it?
- What is the ongoing cost of stewardship?
- What are the safeguards in place to protect?

This, along with a lack of change management frameworks specific to AI Innovation, poses a very real risk to leaders at organizations. This is especially concerning in underserved companies who rely on relevance and agile competitiveness for growth. How can organizations and institutions leverage a reliable and practical framework to assess, adopt, grow and sustain Responsible AI?

AI Innovation: There is a path forward for AI implementation that mitigates failure rate, increases efficiencies, and optimizes revenue. Our "Hoosier AI Maturity Model (HAiMM)" to evaluate an organization's progress towards achieving effective collaboration between humans and AI and measures the benefits across diverse sectors. The model will be more than a checklist. It will integrate both technical and socio-economic dimensions of maturity to foster ethical and uniform use and application of AI. The key components of this model will include: Maturity levels: "AI as a Tool", "AI as an Assistant", "AI as a Partner", "AI as a Catalyst", "AI as an Inspiration" Dimensions: Governance & oversight, Safety, Inclusion, Accountability, Redress This model will include an AI-assisted diagnostic tool using machine learning algorithms and scenario based inputs to assess the AI-adoption maturity and provide actionable recommendations. It will use different ways of thinking, like imagining the future with people, planning how rules might change, and understanding who is involved. This will help make sure things can adapt and react to new ideas about what's right and wrong. It ensures that the principles of responsible AI (RAI) are critically applied throughout each stage of this journey.

Translational Evidence: Empirical evidence from different groups and teams shows that, the more AI is used in a collaborative setting, the more creative and effective the output. Having a maturity model or adoption framework, will help organizations determine where they are and how to move further. It makes for a clear decision tree that has allowances for a variety of use cases and the roadmap will assist them with taking the next steps to make a real impact in achieving Human - AI interoperability. RISE Dimensions: The whole concept behind this is to make sure the AI adoption across Indiana is consistent, responsible, and ethical. This enables secure and inclusive adoption of AI and promotes a reliable and safe journey. The development of the framework will adopt the R.I.S.E. framework. It will be developed through co-design workshops, with safety protocols and informed by cross disciplinary ethics.