Lucy Family
Institute for
Data & Society

Selected Oral & Panel 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

Track: Foundation Models: AI for Science Advances 1

#41: Leveraging Transfer Learning and Molecular Simulations for Accelerated Polymer Property Prediction and Discovery . 1
#43: Rethinking Evaluation in Compound Potency Prediction 1
#66: Are (M)LLMs all you need? A power-aware benchmark of Multimodal Large Language Models for image classification. 2
#177: Leveraging Bioinspired Metaheuristics for Transparent and Reliable Parameter Estimation in Nonlinear Dynamic Models 3
#178: The Emergence of "AI Scientist" systems and the Future of Scientific Innovation 4
#21: Selective Label Smoothing for Predicting Under-Represented Data Samples 4
#126: Multi-Objective Active Learning for Discovering Thermally Conductive and Mechanically Compliant Polymers 5
#176: Active learning (AL) methods for activity coefficient prediction and solid-liquid equilibrium SLE phase diagram construction of Deep Eutectic Solvent systems 6

Track: Human-Centered Responsible AI 7

#3: The case for Search Plurality 7
#29: VOICE: AI-Powered Communication Gateway for Profound Intellectual and Multiple Disabilities 8
#53: The Large Language Model & Community Writing 10
#72: The Human-AI Trust Lab: Advancing Teamwork in the Age of AI 11
#104: Harm as a Design Consideration?: Exploring UX Practitioners' Interpretations and Mitigation Strategies 12
#54: Attention is Not All you Need 13
#83: Toward a Human-Centered Evaluation Framework for Trustworthy LLM-powered GUI Agents 14
#112: In Attaining the Unattainable: Agency and Creativity in AI-generated Liminal Space 16
#121: Ethically And Efficiently Leveraging AI To Effectuate Employment Hiring Decisions 16

Track: Reimagining Global Governance and Policymaking 18

#2: A Model of Algorithmic Wage Determination 18
#25: Distilling heterogeneous treatment effects: Stable subgroup estimation in causal inference 18
#35: Can an Impersonal and Objective Algorithm Become a Better Law Enforcer? Insights from Confucian Criticism of Chinese Legal Philosophy and Practice 19
#36: Building AI Text Classifiers with Peacebuilders: A Human-AI Collaboration to Improve Conflict Analysis and Resolution 19
#38: The Perks and Perils of Machine Learning in Business and Economic Research 20
#52: Assessing Science Use in AI Policymaking: A Case Study of California 21
#117: The Impact of Artificial Intelligence on Labor Markets: A Forecasting Model for Significant Economic Events 22
#128: Stochastic Justice: Inherent Non-Deterministic AI Inference in Legal Decision-Making 22
#138: Unpacking the "Good" in UN AI for Good Projects Using Computational Grounded Theory 23
#174: AI-Driven Detection of Fake News through Emotional Representation Analysis 25

Track: AI for Safety and Safety of AI 26

#37: On Implicit Social Biases in LLMs' Chain-of-Thought Reasoning 26
#39: Data-centric AI for the Early Detection of Suicide Risk in Chilean Schools 27
#98: On Fairness Vulnerability: A Data Poisoning Attack Perspective 27
#101: The Challenge of Privacy in Assessment Data 27
#102: Utilizing AI to Enhance Law Enforcement Response and Public Safety in Underserved Communities 28
#119: Misguided by Machines and Minds: How AI Exploits Our Need to Belong and Fuels Misinformation 29
#131: Implementing Artificial Intelligence in the Hospitality Sector to Combat Trafficking of Children 29

Track: Education and Workforce of the Future ... 31

#7: Advancing AI Capabilities and Evolving Labor Outcomes ... 31
#18: Increasing AI Literacy and Appropriate Use in Nursing Education ... 31
#23: Comparing Statistical and Machine Learning Methods in Missing Data Imputation in Graded Response Model ... 33
#44: AI-Enhanced Reflective Practice: Increasing Access to Faculty Development ... 34
#64: AI-powered, Targeted Instructional Support for Early Childhood Teachers ... 35
#71: Teacher Simulator: Synthetic Dialogue Generation for AI-Powered Instructor Development ... 36
#65: Balancing between Efficiency and Inefficiency: When Economic Sanctions Meet AI/ML ... 37
#95: AI and the Future of Executive Leadership: Evidence-Based Insights on Decision-Making, Skills Development and Workforce Engagement from Executive MBA Students in Pan African Business School ... 37
#157: Bridging the LLM Accessibility Divide? Performance, Fairness, and Cost of Closed versus Open LLMs for Automated Essay Scoring ... 40
#169: Beyond Model-Centric Interpretations: A Framework for MAcro-eXplainability in Generative AI Systems (MAX AI) .. 42

Track: LLMs4All: Large Language Models for Research and Applications in Academic Disciplines ... 44

#172: Emergency Algorithms: RISE-compatible Architectures for Times of Crisis ... 44

Track: Foundation Models: AI for Science Advances

#41: Leveraging Transfer Learning and Molecular Simulations for Accelerated Polymer Property Prediction and Discovery

Authors: Sobin Alosious, Meng Jiang and Tengfei Luo

Abstract: The vast chemical diversity of polymers offers immense potential for developing materials with tailored properties across electronic, electrical, mechanical, structural, and thermal domains. However, traditional experimental and computational methods often face limitations in scalability, predictive accuracy, and data efficiency. To address these challenges, we introduce a multi-task deep learning framework leveraging Graph Neural Networks (GNNs), which effectively captures the complex interplay between polymer structure and multiple functional properties. Our approach represents polymer structures as graphs generated from simplified molecular input line entry system (SMILES) strings, encoding both connectivity and the periodic nature of polymer chains. Our model integrates data derived from molecular dynamics (MD) simulations, density functional theory (DFT) calculations, and experimental datasets. By simultaneously predicting diverse polymer properties such as electronic characteristics, mechanical strength, and thermal performance, our method exploits intrinsic correlations among these properties, resulting in enhanced predictive accuracy and significant improvements in data utilization efficiency.

This facilitates the accelerated discovery of polymers with specifically tailored characteristics, significantly reducing experimental trial-and-error. Importantly, our research demonstrates substantial translational potential. For instance, the proposed model can rapidly guide material design in industries such as flexible electronics, aerospace, automotive manufacturing, and thermal management systems, contributing to improved product performance, reduced material costs, and enhanced sustainability. The framework's inherent scalability and efficiency also democratize access to cutting-edge material informatics tools, empowering researchers and industries in underserved regions and communities that traditionally have limited resources for extensive experimental research. Aligned with Responsibility, Inclusion, Safety, and Ethics (RISE) dimensions, our methodology emphasizes model interpretability and transparency through advanced explainability tools, thus minimizing biases and promoting responsible AI usage. By explicitly addressing transparency and interpretability, we ensure our findings and methodologies can be widely understood, trusted, and adopted by diverse stakeholders, including those from resource-limited research communities. In conclusion, our multi-task GNN-based deep learning framework, informed by MD, DFT, and experimental data, significantly advances polymer informatics, enabling efficient prediction and discovery of multifunctional polymers, fostering socio-economic benefits, and promoting equitable and ethical innovation.

#43: Rethinking Evaluation in Compound Potency Prediction

Authors: Brenda Nogueira, Nuno Moniz, Connor Coley and Nitesh Chawla

Regression tasks play a critical role across many fields, including chemistry, where predictive models help prioritize chemical compounds for experimental testing. In drug discovery, for example, researchers often aim to maximize specific properties like potency—how effectively a compound produces its intended biological effect. However, conventional evaluation metrics and loss functions used to train these models typically optimize for average performance, assuming all prediction values are equally important. This assumption does not align with real-world needs, where certain outcomes are far more valuable than others. In this paper, we highlight the urgent need to revisit standard evaluation and optimization practices in drug discovery and related domains where domain relevance varies significantly. We analyze ten potency classes and compare the outcomes of models optimized using traditional loss functions versus those trained with two domain-sensitive approaches: a feature space design method and a recently proposed loss function that incorporates non-uniform domain preferences. Our results show that models using these domain-sensitive methods consistently identify more unique and higher-performing compounds. These models not only recover compounds found by traditional methods but also prioritize those with greater practical significance. By accounting for domain relevance, these approaches enhance predictive accuracy in the most critical cases and better differentiate between more and less important predictions. This work directly supports the RISE dimensions by promoting responsible and safe AI practices

in high-stakes scientific applications. From a responsibility perspective, aligning model evaluation with real-world priorities reduces the risk of misdirected experimental efforts and increases the likelihood of discovering impactful compounds. By improving model reliability where it matters most, our approach strengthens the safety of AI-assisted decision-making in domains like drug discovery, where errors can have significant downstream consequences. In conclusion, we call on the AI and scientific communities to rethink how regression models are evaluated and optimized. Embedding domain relevance into these processes leads to more effective, ethical, and context-aware AI systems. We invite others to build on this perspective to advance responsible AI in critical fields.

#66: Are (M)LLMs all you need? A power-aware benchmark of Multimodal Large Language Models for image classification.

Authors: Grigorii Khvatskii, Yong Suk Lee, Nuno Moniz, Corey Angst, Maria Gibbs, Robert Landers, Donald Brower, Anna Sokol, Brenda Cruz Nogueira, Satyapragnya Kar and Nitesh V. Chawla

Abstract: Recent research efforts and resources have increasingly been directed toward enhancing transformer-based language models, particularly multimodal language models (MLLMs). However, there has been limited examination of how these sophisticated models perform on relatively straightforward machine learning tasks, such as image classification, compared to traditional, simpler methods. This gap is significant, considering the potential benefits of using simpler models for many practical applications. While extensive work has been done to boost performance metrics like benchmark scores and human evaluations, little attention has been paid to the power consumption of these models during inference. Although some studies have aimed at reducing power use during training, inference-related energy efficiency remains largely overlooked. Improving energy efficiency is critical for sustainable and responsible AI development. Moreover, deploying smaller and more efficient AI models on edge devices offers considerable benefits beyond just energy efficiency. For instance, edge-based AI can greatly improve responsiveness and user experience while simultaneously reducing safety privacy risks and data leaks.

Neglecting these considerations has already led to cases where large, power-intensive models with unverified performance advantages are employed for simple tasks. Such irresponsible deployments could unnecessarily increase greenhouse gas emissions and potentially introduce unexpected performance issues. In many scenarios, advanced multimodal language models may be excessive, and simpler, more efficient models could fulfill the task requirements better. Thus, it is essential to match the complexity and power consumption of AI models to the specific demands of the intended applications. This highlights the importance of a clear, comprehensive approach to measuring both the performance and energy use of various AI models during inference. Such an approach can ensure AI development prioritizes not just raw performance, but also environmental responsibility and practical societal benefits. In our work, we present a novel benchmarking methodology designed to identify the optimal machine learning methods under diverse constraints, including different training and inference budgets as well as power limitations.

Our benchmark is adaptable, capable of evaluating a broad range of machine learning techniques from simple approaches like decision trees to complex frontier multimodal large language models. Although our benchmark does not depend on any particular method for power measurement, we propose using power measurement hardware available on most Intel CPUs. This approach ensures greater accessibility for researchers aiming to benchmark their models. Our research is expected to offer several direct and indirect societal benefits. First, we aim to clearly quantify and communicate the power usage of advanced transformer-based models, both during training and inference. We will also establish baseline measurements (both power and performance) using simpler, classical machine learning methods. These results should encourage developers and engineers to emphasize energy efficiency and carefully assess environmental impacts before deploying models. Additionally, we believe our research will inform future AI model development, helping optimize models not just for performance but also for reduced power consumption.

Finally, by providing accessible benchmarks and promoting efficient models, we hope to accelerate the adoption of edge-based AI solutions, enhancing user experience, responsiveness, and data privacy in daily-use devices. In summary, this research project aligns strongly with the Responsibility and Safety (RISE) dimensions by addressing the energy consumption and environmental impact associated with AI model deployment. By evaluating model performance and identifying gaps, our research provides clearer guidance for safe and responsible use of AI technologies, particularly in performance-critical and environmentally sensitive applications.

#177: Leveraging Bioinspired Metaheuristics for Transparent and Reliable Parameter Estimation in Nonlinear Dynamic Models

Authors: Juan Tejada and Daniel Sanin-Villa

Problem Statement: Accurate parameter estimation in dynamic systems is a longstanding challenge in control engineering, particularly in nonlinear, underactuated systems such as the inverted pendulum. These systems, which serve as proxies for applications in robotics, aerospace, and rehabilitation, are sensitive to uncertainties in physical parameters like mass, inertia, and center of mass. Conventional estimation techniques often fail under real-world conditions due to noise, model mismatches, and non-convexity of the optimization landscape. This limits the deployment of intelligent control strategies in low-cost platforms and safety-critical environments. Addressing this gap requires novel methodologies that are robust, flexible, and accessible.

AI Innovation: Bioinspired algorithms have emerged as powerful tools for parameter estimation in nonlinear problems due to their ability to explore complex search spaces without the need for gradients or convexity assumptions. These methods, inspired by natural behaviors such as evolution (genetic algorithms), collective movement (particle swarm optimization), or cosmological metaphors (multiverse optimization), can adapt to multimodal and highly nonlinear objective functions. Their robustness to local minima makes them suitable for systems with uncertain relationships among variables or partially known physical models. In this study, we present a comparative framework applying Particle Swarm Optimization (PSO), Continuous Genetic Algorithm (CGA), and Salp Swarm Algorithm (SSA) to estimate physical parameters of a double inverted pendulum. The approach leverages torque-based fitness evaluation, dynamic modeling via the Euler–Lagrange formalism, and extensive statistical validation over 1000 simulation runs. The SSA achieved the lowest root mean square error (RMSE = 0.015 N·m), outperforming the other algorithms in both accuracy and consistency.

Translational Evidence: The proposed estimation strategy was implemented and validated on a physical pendulum prototype, demonstrating its viability for real-time deployment in educational, industrial, and assistive robotics contexts. The results show high fidelity between predicted and experimental torque values, enabling more effective control policies such as PID, fuzzy logic, or model predictive control. Moreover, the metaheuristic-based estimation pipeline is hardware-agnostic and requires minimal system-specific tuning, making it especially attractive for low-resource settings. The public availability of the model and code supports replication and adaptation. Its use in educational robotics, for example, empowers institutions in underrepresented regions to teach control system design with tools that rival high-end commercial simulators.

RISE Dimensions Responsibility: This work adheres to responsible AI principles by integrating interpretable models grounded in physical laws, avoiding black-box behavior, and enabling transparent debugging and verification of system performance. Inclusion: By applying AI to low-cost, open-source mechatronic platforms, the methodology facilitates technological transfer to underserved educational and industrial communities, reducing the digital divide. Safety: Improved parameter estimation translates into enhanced controller stability and responsiveness, which is critical in applications involving physical human–robot interaction or autonomous navigation in uncertain environments. Ethics: The openness of the data, algorithms, and validation process supports ethical reproducibility and trust. The algorithms used avoid discriminatory bias and can be deployed in equitable public systems. This work embodies an interdisciplinary integration of artificial intelligence, mechanical modeling, control theory, and engineering education to produce socially relevant, ethically grounded, and technically robust solutions for real-world dynamic systems.

#178: The Emergence of “AI Scientist” systems and the Future of Scientific Innovation

Authors: Marko Grobelnik

Abstract: The fairly recent appearance of “AI Scientists” is opening more questions than answers. “AI Scientists” are systems, started appearing mainly in 2025, which automate the scientific process from start to finish, either with partial significant human involvement at times or with no human involvement at all. This includes all the key phases of the scientific process from hypothesis generation, system building (e.g., programming), evaluation, correction/improvement, up to scientific paper writing or grant proposal writing. As of 2025, there are several such (mostly) prototypes in operation, with good prospects that this line of development could soon reach much more mature levels where humans will increasingly be out of the loop in the creative parts of the scientific process. These developments are in some ways surprising, especially considering how quickly they appeared after the relatively recent public availability of GenAI systems in 2022/2023. The assumption that “AI Scientist” systems will evolve in the near future raises many questions: philosophical, ethical, societal, funding-related, business, organizational, educational, and more. In our contribution, we will discuss the nature of human knowledge as we know it (mostly incremental with occasional breakthroughs), and how automating the process of inventions can change and accelerate the processes and broader scientific ecosystem. The relevant question is if automating science with the current and expected future AI technology has a convergence point or if it may diverge in yet unknown directions. This could easily lead towards automated inventions built on top of previously automated inventions in various fields of science. Also, in the context of the soon expected significant transition of AI into the physical world with more or less autonomous robots, where science is still dominated by human work in terms of experimentation and deployment. We will address various societal challenges related to automating science, some of which are already observable in 2025. This includes at least two main levels: (a) the geopolitical race to maintain or gain strategic advantages in a politically divided world, and (b) the more traditional competition between companies to dominate the markets and maximize profits. An important factor to consider is the noticeable decrease in the time of the traditional scientific cycle, from hypothesis to invention and then to market-ready product. The key question here is: what will the “new normal” be, and where will the ongoing transformational times stabilize in the presence of seemingly exponential developments of AI in recent years?

#21: Selective Label Smoothing for Predicting Under-Represented Data Samples

Authors: Doheon Han, Nuno Moniz and Nitesh Chawla

Problem Statement: The imbalance issue is still problematic since it harms model performance in machine learning. Above all, the imbalance exists in most real-world problems, leading to worse decision making. In general, the imbalance refers to a class imbalance issue which makes a model biased toward predicting majority samples. However, the imbalance can happen among individual samples regardless of class distribution, and this work suggests one solution for the problem. AI Innovation Labels for data samples can be regarded as goals for each sample when Binary Cross Entropy (BCE) loss function is used in neural networks, since BCE measures error by calculating the differences between the probability, output of a sample, and the label. We can assign different weights to each sample by adjusting the labels, and this can mitigate the imbalance issue. The Selective Label Smoothing (SLS) technique applies label smoothing to only under-represented samples to improve overall model performance. Translational Evidence SLS can give more or less weights to any selected group of samples when a model is learning, therefore, we can care more about the under-represented people or effects when making decisions. RISE Dimensions We have witnessed many cases of fairness issues caused by imbalance, and this study suggests one ready-to-use method for the issue, and is expected to be widely used for diverse tasks as long as it needs binary classification.

#126: Multi-Objective Active Learning for Discovering Thermally Conductive and Mechanically Compliant Polymers

Authors: Yuhan Liu, Jiaxin Xu, Renzheng Zhang and Tengfei Luo

Abstract: Polymers are extensively used in industrial applications such as flexible electronic devices and thermal interface materials, due to their mechanical flexibility, corrosion resistance, lightweight, and low cost. However, conventional polymers typically exhibit poor heat transfer capabilities, significantly limiting their application in high-performance thermal management. Therefore, identifying amorphous polymers that possess high intrinsic thermal conductivity (TC) while maintaining desirable mechanical flexibility is of great significance. Traditionally, new polymeric materials have been designed based on experimental intuition, followed by trial-and-error synthesis. Due to inconsistent synthesis processes and characterization methods, obtaining reliable pure-polymer data often involves long testing periods and high costs, limiting the efficiency of large-scale polymer screening. Recent advances in materials informatics, particularly through machine learning, have emerged as promising approaches to accelerate materials discovery. However, most polymer design efforts focus on optimizing a single property, while simultaneously optimizing multiple, sometimes competing properties remains a major challenge.

In this work, we develop an active learning (AL) framework based on multi-objective Bayesian optimization (MOBO) to efficiently identify a set of optimal polymers with high TC (κ) and low bulk modulus (B). We first curate an initial small dataset by selecting 87 polymers from various polymer classes in the PoLyInfo database and calculating their TC and bulk modulus via high-throughput MD simulations. To consider the influence of polymer morphology on computed properties, MD simulations are performed using three distinct initial structures for each polymer, and their averaged TC and bulk modulus values are used as the training dataset. Polymer structures are represented using polymer embedding, and then independent Gaussian Process Regression (GPR) surrogate models are constructed for both properties (κ and B) to capture structure-property relationships. These models are used to explore a large pool of thousands of unlabeled polymers, where predicted means and uncertainties guide the acquisition function in MOBO, recommending new candidates for further MD labeling and evaluation. The newly selected polymers are iteratively integrated into the training dataset, updating the GPR models to initiate a new iteration until a set of Pareto-optimal polymers with high κ and low B is obtained. To achieve efficient multi-objective optimization, we employ the q-noisy expected hypervolume improvement (qNEHVI) algorithm as the acquisition function. The AL workflow terminates once 20 polymers with TC above and bulk modulus below predefined threshold values are identified.

Additionally, we evaluate the synthesizability of the optimal polymers by calculating their synthetic accessibility scores and further provide insights into the intrinsic relationships between polymer microstructures and their properties. This AL-driven MOBO framework effectively accelerates the discovery of multifunctional polymers, reducing the traditionally prolonged and costly development cycles from decades to months. The identified polymers provide sustainable and lightweight alternatives to conventional metal-based heat dissipation components. Ensuring responsible AI practices, we validate computational simulations against experimental benchmarks to guarantee accuracy and reliability. We further uphold ethical standards through open-source data sharing and collaborative research practices. By efficiently transitioning computational predictions into experimental validation, this scalable strategy facilitates the design of next-generation polymer materials and closely aligns with RISE dimensions.

#176: Active learning (AL) methods for activity coefficient prediction and solid-liquid equilibrium SLE phase diagram construction of Deep Eutectic Solvent systems

Authors: Paul Amagada, Edward Maginn and Yamil Colon

The future of energy storage depends on safer, more efficient technologies, yet current flow battery electrolytes face challenges, including poor electrochemical stability and safety risks. My research explores deep eutectic solvents (DESs) as next-generation electrolytes for redox flow batteries (RFBs), offering advantages like low volatility, non-flammability, and environmental friendliness. However, a key gap in the literature is the lack of systematic methods for predicting SLE phase behavior, which is crucial for designing stable electrolytes. My research addresses this gap by developing an active learning-driven framework to optimize SLE phase diagram construction, enabling efficient screening of DES-based electrolytes to improve flow battery performance and support scalable energy storage solutions.

Understanding SLE phase diagrams is crucial for designing DESs for energy storage. However, constructing these diagrams is time-consuming and error-prone due to the need for multiple experimental iterations. To address this, I will develop an AL framework that efficiently selects informative data points, reducing the experimental workload while improving predictive accuracy. This study will initially focus on Type III and Type V DESs, including choline chloride/urea, thymol/menthol, and thymol/trioctylphosphine oxide (TOPO), before extending to more complex mixtures. The first step involves selecting an initial data point at the estimated eutectic composition, followed by generating virtual data points across the temperature range of the DES mixtures to establish a foundation for modeling. A Gaussian Process (GP) model will be trained using these data points to predict activity coefficients. GP models are particularly suited for this study due to their ability to quantify prediction uncertainties, a critical feature for AL. By assuming solid-phase immiscibility, the SLE curve for a component in a DES mixture is given by the equation: $log\ log(x\_i\ y\_i\ ) = (\Delta mh\_i)/R(1/T - 1/T\_(m,i)\ )$, where x_i and y_i represents the liquid-phase composition and activity coefficient of component i, while T , and Δmh correspond to its melting temperature and enthalpy, respectively. T is the absolute temperature of the system, and R is the gas constant. The GP model will compute the predicted SLE curve and identify a possible eutectic point, iteratively refining its accuracy using AL techniques.

AL is a machine learning strategy that enables an algorithm to proactively identify the most informative training data points to improve its performance. To optimize the SLE data collection process, an acquisition function will be computed to assess the uncertainty and informativeness of all possible SLE data points. The next experimental measurement will be selected where the acquisition function is maximized, ensuring that each data point provides maximum value toward refining the phase diagram. The developed AL algorithm in this research will drastically reduce the number of experimental measurements required for SLE phase diagram construction, while maintaining high predictive accuracy. The AL framework will also facilitate the large-scale screening of potential DES systems, enabling the rapid identification of optimal compositions for energy storage applications. This research also aligns with the RISE dimensions by promoting responsibility, inclusion, and ethics in the development of energy storage technologies.

The use of deep eutectic solvents (DESs) as electrolytes for redox flow batteries offers a safer and environmentally friendly alternative to conventional electrolytes, reducing the risks associated with flammability and toxicity while improving electrochemical stability. This approach contributes to responsible innovation by ensuring the technology is both scalable and sustainable. It allows for inclusion by enhancing access to efficient energy storage solutions, particularly in underserved regions, where traditional energy infrastructure may be lacking. Moreover, the research is grounded in ethical principles, ensuring that AI-driven advancements in materials science are deployed to solve critical societal challenges without compromising safety, fairness, or environmental sustainability. By addressing both technological and societal needs, this work exemplifies how AI can be used to create positive, long-lasting impact in energy storage and beyond.

Track: Human-Centered Responsible AI

#3: The case for Search Plurality

Authors: Shiran Dudy

Abstract: In light of Phillips' argument regarding the impracticality of Search Neutrality~\cite{phillips2023algorithmic}, which highlights the influence of non-epistemic factors in shaping result prioritization, this work aims to engage with this limitation by critically examining current design practices in search engines. We contend that the prevailing emphasis on prioritization and the rigid hierarchical ordering of ranked lists deserve closer scrutiny. To address these gaps, we propose \textit{Search Plurality}—a reimagined information-seeking interface centered on enhancing information accessibility and reorganizing data. We argue that this approach, particularly for specific search intents, fosters a more inclusive user experience, promotes deeper topic comprehension, and provides an avenue to mitigate echo chambers. The common assumption is that search engines deliver results (or items) most relevant to the queries we enter. However, relevance exists in a multi-dimensional space, making it impossible to directly compare items across different dimensions. Despite this, the ordered presentation of items suggests that such comparisons have occurred, and therefore Phillips argues that the ordering ought to be shaped by non-epistemic ideologies – deeming search neutrality impossible. These influences may introduce biases that affect the manner through which we acquire knowledge and the kinds of knowledge acquired through these tools.

Ranked lists have been a foundational feature of search engines since their inception. While effective, this design has notable limitations. First, ranking enforces a hierarchy of items even when one may not naturally exist, as items with numerically similar scores are imposed an artificial order of importance. Second, a study by Chun et al.~\cite{chun2022power} indicates that users rarely explore all the items offered, limiting our attention to focus on a few items, and by that narrowing our understanding the depth or breadth of possibilities of topics (irrespective of the usefulness of an abundance display). Third, the current presentation of items requires users to individually make sense of them, as there is no organizational structure beyond their ranking to indicate a justification for their inclusion or how they relate to the query. Fourth, the selected items often align with what the majority of users find relevant~\cite{munton2022answering}, thereby excluding minority perspectives and experiences from these lists.

We introduce the term Search Plurality to highlight an approach that emphasizes the diverse ways a query can be explored. While this method does not entirely eliminate the impartiality issues previously discussed, it aims to mitigate their impact by prioritizing diversity in addressing queries and removing the traditional, hierarchical ordering of search results. This concept introduces a design that emphasizes category-based organization, using blurbs to convey the scope and structure of a query. Categories highlight ambiguity, diverse perspectives, or multifaceted aspects of a topic, encouraging users to explore independently. An unordered display of categories reveals the query's breadth, while clicking on a category uncovers its depth through relevant links. Users can choose from multiple interfaces, gaining insights within a label-organized framework. By adopting rich discovery interfaces, epistemic environments can prioritize information accessibility, validate minority perspectives, and reduce polarization, fostering healthier and more informed societies.

#29: VOICE: AI-Powered Communication Gateway for Profound Intellectual and Multiple Disabilities

Authors: Jarek Nabrzyski, Michal Kosiedowski, Ed Fennell and Kristin Wier

Abstract: This paper presents VOICE, a groundbreaking initiative that aims to revolutionize communication for individuals with Profound Intellectual and Multiple Disabilities (PIMD) through multimodal AI-driven assistive technologies. PIMD individuals often face extreme barriers in both verbal and non-verbal communication, leading to significant isolation and limited engagement with caregivers, family members, and society. Despite these challenges, research and innovation efforts for PIMD individuals remain minimal, leaving caregivers with few effective solutions.

VOICE centers on the development of an advanced AI-based communication system, uniquely designed to bridge the communication gap for those who face significant challenges in verbal and non-verbal expression. Unlike traditional assistive technologies that rely on text-based or motor-dependent inputs, VOICE leverages multimodal AI to interpret and translate a range of non-verbal cues into actionable communication signals. These include facial expressions, body movements, vocalizations and physiological signals, as well as eye tracking and neurological biomarkers. By utilizing computer vision, natural language processing, and real-time signal processing, VOICE will empower individuals with PIMD to express emotions, needs, and intentions more effectively. This shift in approach moves beyond diagnosing limitations and towards enhancing interaction and independence.

Model for University-Care Center Collaboration: Logan Center and University of Notre Dame, and Poznan Supercomputing and Networking Center and PPIMD Centers in Europe. A key element of VOICE is its collaborative research and development model, which fosters partnerships between universities and care centers. The Logan Center, a leading institution in autism and PIMD care, and the University of Notre Dame, through its Center for Research Computing, serve as a model for how interdisciplinary collaboration can drive innovation in assistive AI technologies.

Logan Center provides real-world environments for AI testing and patient interaction, while the University of Notre Dame contributes expertise in AI, cloud computing, and data security. Caregivers and healthcare professionals work closely with researchers to train AI models that adapt to individual patient needs. Patients and caregivers provide direct feedback, enabling the refinement of AI interpretations. This iterative process ensures that the AI system is not only scientifically robust but also practical and user-friendly for caregivers and PIMD individuals.

The collaboration also ensures that the technology is deployable in both home and clinical settings. The edge-computing and cloud-based infrastructure facilitates scalable deployment to care centers globally. Through this university-care center model, VOICE establishes a sustainable framework that can be expanded to other institutions, integrating real-world caregiving expertise with state-of-the-art AI research. CRC also collaborates with institutions similar to Logan located in Japan and Europe.

A similar model exists in Poland, where Poznań Supercomputing and Networking Center (collaborator on VOICE project) collaborates with Home for Children and Youth with Disabilities in Poznań, run by the Sisters of Seraphim. This partnership follows the same principles of interdisciplinary collaboration, leveraging AI expertise alongside direct caregiver engagement to create and refine assistive communication technologies for PIMD individuals. The success of these models highlights the potential for scaling such initiatives internationally, integrating AI research with hands-on caregiving expertise.

Innovation and Impact: VOICE addresses critical challenges faced by PIMD individuals by facilitating non-verbal communication and developing cognitive biomarkers to track progress in therapy and intervention. It creates scalable, AI-driven solutions that can be deployed in diverse care environments, bridging the research gap by shifting the focus from disabilities to preserved cognitive abilities. The project integrates multiple AI-driven components to interpret non-verbal cues in real-time. Sensor-based AI systems capture physiological and environmental signals, while machine learning for gesture and expression recognition translates body movements and facial expressions into actionable insights. Voice and sound analysis help identify meaningful vocal patterns to aid communication, and cloud-connected AI processing ensures accessibility across different care environments. By leveraging AI for multimodal data interpretation, VOICE opens new avenues for communication and interaction among PIMD individuals, caregivers, and society.

A crucial aspect of the VOICE system is the development of a small, low-power edge device for real-time data collection and analysis. This device is designed to operate with power consumption comparable to a smartphone, making it both efficient and accessible. The sensors embedded within this compact and cost-effective system include non-invasive cameras, microphones, an electronic nose, remote heart rate monitors, and temperature sensors. This multimodal AI-powered device enables continuous monitoring without intruding on the patient's daily routine, providing valuable insights into their cognitive and physical states. Additionally, the technology has dual-use applications, extending beyond PIMD to support baby monitoring, post-stroke patient recovery monitoring, and other use cases where real-time physiological and behavioral analysis can enhance care and treatment outcomes.

Student Engagement: A fundamental component of the VOICE initiative is the active engagement of undergraduate students in the research, development, and implementation of the system. Students from computer science and neuroscience disciplines at both the University of Notre Dame and Rensselaer Polytechnic Institute (RPI) play a critical role in advancing the project. The baseline infrastructure of the system is designed to allow students to contribute in multiple ways, including developing different components of the AI models and participating in model training, personalization, and refinement.

This hands-on engagement provides valuable research experience and fosters a new generation of scientists and engineers dedicated to assistive AI technologies. By involving students in the technical and ethical aspects of the project, VOICE not only enhances their academic growth but also ensures that innovative ideas continuously flow into the system. This educational collaboration creates a sustainable ecosystem for research and development, benefiting both academia and the broader community of caregivers and individuals with PIMD.

Conclusion: The VOICE project represents a transformative step toward AI-assisted communication for PIMD individuals, providing a scalable, multimodal, and caregiver-inclusive solution. By fostering deep collaboration between universities and care centers, as exemplified by the Logan Center and the University of Notre Dame, the project not only advances technology but also sets a precedent for ethical and effective AI deployment in the disability support sector. This initiative will redefine what is possible in assistive communication, offering individuals with PIMD a voice where none existed before.

#53: The Large Language Model & Community Writing

Authors: Matthew Kilbane

Problem Statement: This interdisciplinary presentation examines the literary ramifications of large language models (LLMs) through a case study of Lillian-Yvonne Bertram's 2024 poetry collection A Black Story May Contain Sensitive Content. When the online journal Diagram awarded this volume, which was written in collaboration with GPT3, their 2023 Chapbook Award, Bertram's work became a flashpoint in heated conversations about literature and generative AI. Scarcely understood at the time was the degree to which, far from uncritically adopting LLMs, A Black Story May Contain Sensitive Content further develops Bertram's ongoing project of immanent technological critique: this work "collaborates" while espousing all the while a clear sense of the technology's limitations and dangers. Indeed, she foregrounds these dangers at every turn; they constitute a fundamental component of the book's subject matter. Bertram engages with LLMs in order to wrest a socially significant art from the ideological construction of the technology itself. As a Black poet interested in charting both the underlying racial politics of deep learning models as well as the degree to which their own poetry is "haunted" by the pioneering example of earlier Black poets, Bertram writes that their "work fine-tuning large language models is influenced by this kind of haunting, the what if, and inquires into how these models model voices that no longer exist, voices of writers we don't often get to hear, such as Gwendolyn Brooks."

In this presentation I explain both the theoretical background and evident success of this effort, while also grounding my account of one poet's use of LLMs in a longer tradition of poetic collaboration and community writing. In short, it is no surprise that Bertram chooses to fine-tune her language model on a corpus of texts by Gwendolyn Brooks, a prominent twentieth-century poet whose commitment to the social life of poetry involved engagement across her life with a series of pivotal workshops, from Inez Cunningham Stark's Chicago Southside workshops in the 1940s to her own workshops in the late 1960s with the Blackstone Rangers. Positioning the LLM alongside the community-based workshop may seem like a rather perverse category error, but taking my lead from Bertram's work, I demonstrate how this odd juxtaposition furnishes insights into the limitations and possibilities of generative AI for literary practice.

AI Innovation: The project from which this presentation emerges argues that the task of comprehending the cultural impact of generative AI—its influence on everyday practices of reading and writing—demands a methodological framework informed not only by computer science and associated disciplines in the sciences but also by several specific subfields in the humanities long dedicated to the study in transformations of reading and writing: literary history, cultural theory, and the sociology of literature.

Translational Evidence: I plan to publish a peer-reviewed article addressed to scholars at the intersection of critical media studies and literary theory, but this publication will also offer a practical framework, oriented toward teachers of creative writing at the secondary and tertiary levels, for responsibly incorporating generative AI into curricula.

RISE Dimensions: This project aligns with the first two dimensions of RISE. Bertram's pathbreaking poetic practice is fundamentally geared toward the discovery of responsible artistic uses of deep learning models—uses that remain sensitive to the ideological assumptions at play in generative AI's design and deployment. By "translating" these insights for an audience of creative writing teachers at the secondary and tertiary level, we can extend the reach of this responsible comportment toward new literacy technologies. At the same time, Bertram's focus on the specifically Black experience of engaging with generative AI, and her critical concern for the way new technologies of knowledge-production and surveillance can rehearse and deepen patterns of racial discrimination, affiliates this project with the goal of upholding principles of inclusion in the development of artificial intelligence.

#72: The Human-AI Trust Lab: Advancing Teamwork in the Age of AI

Authors: Trenton Ford, Michael Yankoski, Mohammed Almutairi, Charles Chiang, Nandini Banerjee, Matthew Belcher, Tim Weninger and Diego Gomez-Zara

Problem Statement: Human and AI Teams (HATs) have become increasingly prevalent across a wide range of domains, in which AI agents facilitate and collaborate with human team members. The complex relational dynamics that emerge within HATs demand more study and scrutiny as HATs proliferate. How can researchers and simulators employ AI to simulate these HATs with more accuracy and variety? Animating questions our team is pursuing include: RQ1: How might an AI system adapt its communication characteristics for certain personality types in human teammates? RQ2: How might AI systems work to enhance, protect, and repair trust with human teammates before and after trust is harmed? RQ3: When should HAT transparency be increased (ie: when do human teammates require more explanation), and when should transparency be decreased (ie: to reduce cognitive load for human teammates?) RQ4: What does it mean for an AI system to "trust" its human teammates? Should trust be understood as a bi-directional process in HATs?

AI Innovation: The goal of this project is to design, build, and test an AI system that simulates HATs. Based on Large Language Models (LLMs) modules, we design a simulation framework that allows users to create and simulate customizable HATs for further research. While most previous work has focused on conversational tasks [1], [2] or workflow tasks [3], [4] performed by AI agents, we build these simulated agents using state-of-the-art LLMs—which provide several human capabilities, including rationale, memory, and decision-making—in a simulated environment that has physical and temporal properties. As such, agents in our system must learn from their environment, objects, and team members to complete their tasks. We have constructed and experimented with HAT arrangements composed of HDTs (Human Digital Twins) and BOTS (Bounded Operational Teammates). Our HDTs (Human Digital Twins) are highly adapted versions of Comp-HuSim Agents [5], replete with advanced personality, psychographics, and memory systems. BOTs are the HATs AI teammates, whose capacities are restricted to certain domains of knowledge, and who retain memories of the interactions they have had with their HDT team members. By employing a centralized architecture that coordinates the actions, events, and states of the environment, we provide HDTs and BOTs with situational awareness and knowledge to succeed in their tasks. This framework, called PuppeteerLLM, coordinates the actions and enables HDTs to communicate with each other. Our highly configurable and iterable experimentation HAT framework allows us to design and experiment along a significant range of variables and scenarios. Users can see the progress of the simulation through a dynamic and interactive interface, showing a map and log of the agents' actions.

Translational Evidence: As a part of our larger research efforts, we will continue to design, construct, and refine the software libraries and platforms needed to run these simulations and HAT teaming experiments at larger scales. For RISE AI 2025, we are eager to demonstrate and present our system and obtain feedback and insight from other researchers in this space, particularly regarding increased understanding of how these kinds of HAT teaming arrangements may be used to refine AI system adaptation methodologies to human user dynamics and needs. At RISE 2025, we will present no fewer than 5 HATs, each with distinct configurations and members. We will conduct a set of no fewer than 25 experiments for each HAT configuration to run, with each experiment constituting a specific scenario. Lastly, we will conduct post-hoc "subjective" interviews with each HAT member to better understand and interpret their unique perspective of the HAT, in the given experimental scenario.

RISE Dimensions: This project addresses how trust and transparency in HATs can be enhanced (or hindered) by their interactions and training processes. Trust is fundamental for effective collaboration in teams. However, AI agents can lack mechanisms to establish and maintain trust with human team members, and it is even harder to repair trust when a conflict or issue emerges. We will address the different challenges of designing AI agents—including their personalities, communication styles, and teamwork strategies—when transparency and trust are the main objects of study. Furthermore, this project explores the ethical implications for AI trust, exploring whether it should be a bidirectional process and its distinctions. Through different experiments and configurations, this project will examine the consequences of different customizations of trust and transparency, which will alter the cohesion, safety, and performance of these HATs.

References:

1. J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, and P. Liang, "Generative agents: Interactive simulacra of human behavior," Proceedings of the 36th, 2023, [Online].
2. Y. Li, Y. Zhang, and L. Sun, "MetaAgents: Simulating interactions of human behaviors for LLM-based task-oriented coordination via collaborative generative agents," arXiv [cs.AI], Oct. 10, 2023. [Online].
3. Z. Liu et al., "AgentLite: A lightweight library for building and advancing task-oriented LLM agent system," arXiv [cs.MA], Feb. 23, 2024. [Online].
4. C. S. Xia, Y. Deng, S. Dunn, and L. Zhang, "Agentless: Demystifying LLM-based Software Engineering Agents," arXiv [cs.SE], Jul. 01, 2024. [Online].
5. C. Fan, Z. Tariq, N. Saadiq Bhuiyan, M. G. Yankoski, and T. W. Ford, "Comp-HuSim: Persistent digital personality simulation platform," in Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA: ACM, Jun. 2024, pp. 98–101.

#104: Harm as a Design Consideration?: Exploring UX Practitioners' Interpretations and Mitigation Strategies

Authors: Ritika Gairola and Colin M Gray

Abstract: Harm—whether individual or collective, material or non-material—has emerged as a critical lens for understanding the ethical complexity of UX practice. The concept of ethical complexity (Gray & Chivukula, 2019) describes the multifaceted and dynamic challenges UX practitioners encounter when navigating ethical considerations in their work, shaped by mediating factors such as organizational structures, personal values, and external ethical frameworks. It arises from the interplay of competing priorities and knowledge forms (Gray, Gairola, et al., 2024) (e.g., user welfare vs. business goals, legal compliance vs. moral responsibility) along with situational constraints and complexities (Stolterman, 2008) inherent to design practice. When designers fail to address overlapping complexities—like balancing user needs, business goals, and ethical principles—blind spots form, allowing harmful practices like dark patterns to take root. As frontline decision-makers shaping user experiences, designers hold a critical responsibility: their ability to confront complexity directly influences whether ethical risks are mitigated or amplified, particularly when it comes to deceptive design practices, such as dark patterns.

With the rise of AI, these risks are magnified, as AI-enhanced deception introduces new ways to manipulate and influence users. While much research focuses on the immediate effects of AI-enabled deception (Sarkadi et al., 2023), little attention has been paid to the future impact of AI-enhanced deception (Sarkadi, 2021), which could evolve to exploit users in increasingly sophisticated ways, requiring urgent consideration and action. Recently, (Santos et al., 2024) maps the harm caused by manipulative design to legal frameworks for assessing harm, exposing misalignments between ethical intent and regulatory enforcement. This highlights the need for new approaches to harm mitigation in UX design—not as an afterthought, but as an integral part of the design process. Despite the growing recognition of ethical risks, little is known about how UX practitioners interpret, perceive, and respond to these harms in their daily work. How do they navigate the tension between ethical complexity, knowledge constraints, and legal frameworks when harm is a potential outcome?

This study seeks to explore these questions by examining UX practitioners' sensemaking processes, ethical reasoning, and coping strategies to understand how they conceptualize and address harm across different contexts, such as dark patterns, accessibility, and data privacy. We will conduct a series of co-design workshops (Sanders & Stappers, 2008) with UX practitioners, asking participants to engage with two constructive tasks. First, we will create flashcards and scenarios leveraging Santos et al.'s (2024) harm taxonomy to present ethically-nuanced challenges relating to AI and business applications to the participants, giving them a space to consider which harms they would focus on and how they would address these harms in their everyday work. Second, we will allow practitioners to select tasks they face in their actual work environment, using the harm taxonomy to consider potential action plans for future work, inspired by previous co-design efforts (Gray, Obi, et al., 2024). The goal is to understand- "How do UX practitioners perceive and prioritize ethical harms, particularly in relation to AI-enhanced deception, within their design practices?" By grounding Santos et al.'s (2024) harm taxonomy in practitioner insights, this work fosters a sense of accountability in UX practice and aligns

with multiple RISE dimensions. It encourages designers to recognize and address the ethical implications of their decisions, thus promoting responsible design choices that prioritize user welfare. The focus on identifying and addressing ethical harms is inherently linked to creating safer digital environments.

By unpacking how practitioners navigate these tensions, this study helps to enhance user safety through more thoughtful, ethical design decisions that reduce the potential for harm. The study directly addresses ethical considerations by advancing understanding of harm as a critical factor in design. By co-creating strategies to align design decisions with legal and ethical standards, the work emphasizes the ethical responsibility of UX practitioners to safeguard user interests while balancing business needs and technical constraints.

References:

- Gray, C. M., & Chivukula, S. S. (2019). Ethical mediation in UX practice. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Article Paper 178. Link
- Gray, C. M., Gairola, R., Boucaud, N., Hashmi, M., Chivukula, S. S., Menon, A. R., & Duane, J.-N. (2024, July). Legal trouble?: UX practitioners' engagement with law and regulation. Designing Interactive Systems Conference. DIS '24: Designing Interactive Systems Conference, IT University of Copenhagen Denmark. Link
- Gray, C. M., Obi, I., Chivukula, S. S., Li, Z., Carlock, T. V., Will, M. S., Pivonka, A. C., Johns, J., Rigsbee, B., Menon, A. R., & Bharadwaj, A. (2024). Building an ethics-focused action plan: Roles, process moves, and trajectories. Proceedings of the CHI Conference on Human Factors in Computing Systems, 116, 1–17. Link
- Santos, C., Morozovaite, V., & De Conca, S. (2024). No harm no foul: how harms caused by dark patterns are conceptualised and tackled under EU data protection, consumer and competition laws. Link
- Sarkadi, S. (2021). Deception [University of London, King's College]. PDF
- Sarkadi, S., Mei, P., & Awad, E. (2023). Should my agent lie for me? A study on attitudes of US-based participants towards deceptive AI in selected future-of-work. Adaptive Agents and Multi-Agent Systems, 345–354. Link
- Stolterman, E. (2008). The nature of design practice and implications for interaction design research. International Journal of Design, 2(1), 55–65. Link

#54: Attention is Not All you Need

Authors: Marianna Ganapini, Giuseppe Riva, Enrico Panai and Massimo Chiriatti

Abstract: The rise of artificial intelligence as an extension of human cognition marks a paradigm shift in how we process and engage with information. This paper introduces "System 0," a novel, AI-based cognitive system in which AI operates as an external augmentation of human thought, complementing Kahneman's Systems 1 and 2. The goal of the paper is to show the transformation of the attention economy under the influence of System 0 and to propose a practical framework for organizations to navigate this transformation.

By analyzing the integration of AI into cognitive processes, we illustrate how System 0 is evolving from a passive tool for information retrieval into an active mediator of human decision-making—fundamentally reshaping attention, perception, and monetization in the digital economy. We define System 0 through three key characteristics: its epistemological role in structuring information for human cognition, its computational foundation as a pre-semantic processor, and its unique structural hybridity, blending elements of both intuitive and analytical reasoning. This cognitive evolution has profound implications for brand-consumer interactions, necessitating a shift toward AI Optimization (AIO) as a new strategic paradigm for business success.

A central focus of this paper is the transformation of the attention economy under the influence of System 0. Traditionally, the attention economy has been driven by human cognitive limitations—advertisers and platforms competed for finite human attention spans by leveraging persuasive design, behavioral nudges, and emotionally engaging content. However, as AI increasingly intermediates human attention, it is no longer just a resource to be captured but a dynamic process shaped by algorithmic filtering, recommendation systems, and AI-driven personalization. System 0 introduces a new form of attention arbitration, where AI selectively structures and curates information before it even

reaches conscious human awareness. This shift reconfigures how businesses, advertisers, and content creators must approach engagement, requiring them to optimize not just for human cognition but also for AI-mediated perception.

Building on the extended mind theory and advancements in AI capabilities, we propose a practical framework for organizations to navigate this transformation, centered on four critical steps: structuring data for AI parsing, real-time performance tracking, mastering contextual relevance, and developing AI personas. In this evolving landscape, success is no longer determined solely by capturing human attention but by understanding how AI itself prioritizes, processes, and amplifies information. Our analysis suggests that thriving in the AI-driven attention economy requires organizations to engage not only with human users but also with AI systems, recognizing their pivotal role in shaping how value is created and captured in digital environments.

#83: Toward a Human-Centered Evaluation Framework for Trustworthy LLM-powered GUI Agents

Authors: Chaoran Chen, Zhiping Zhang, Ibrahim Khalilov, Bingcan Guo, Simret Gebreegziabher, Yanfang Ye, Ziang Xiao, Yaxing Yao, Tianshi Li and Toby Jia-Jun Li

Abstract: The rise of large language models (LLMs) has transformed Graphical User Interface (GUI) automation across web applications, mobile devices, and operating systems. Traditional automation frameworks, such as Selenium, rely on static action scripts and predefined rules to automate workflows. While effective in structured tasks, these tools lack flexibility and require manual scripting, making them struggle with dynamic, context-sensitive interfaces. Recent advancements in LLMs have led to the development of LLM-powered GUI agents, offering unique capabilities to overcome these challenges. A GUI agent is an autonomous system that perceives and interprets UI elements by analyzing screenshots or GUI source files, translates user commands into sequential actions using LLMs, and interacts with GUIs by executing actions such as clicking, typing, and tapping. Unlike traditional GUI automation, GUI agents further enhance automation by interpreting natural language commands, processing multimodal content, and dynamically simulating user actions.

For example, OpenAI's Operator and Claude's Computer Use assist users in completing complex web forms and navigating dynamic websites based on high-level instructions, eliminating the need for pre-programmed scripts. Beyond automation, GUI agents improve accessibility by allowing non-technical users to interact with applications using natural language prompts. Integrated with assistive technologies like screen readers, they enhance usability for individuals with disabilities. From price comparisons to automated email responses, GUI agents extend automation beyond software testing to everyday tasks, making technology more accessible and improving productivity.

Privacy and Security Risks in GUI Agents: As GUI agents advance, privacy concerns have emerged, affecting user trust. Research shows that even commercial models like GPT-4 struggle with privacy reasoning, sometimes exposing sensitive information. Building on prior work, we categorize GUI agents' privacy and security risks into three key areas: (1) amplified data leaks due to direct access to sensitive data and frequent third-party interactions, (2) diminished user control over privacy and security, and (3) insufficient guardrails against data breaches and adversarial attacks.

Amplified Data Leaks: GUI agents often require access to sensitive user data. Unlike direct LLM prompting, where users can redact personal details, GUI agents need unfiltered information to complete tasks. For example, booking a flight requires the agent to process travel details, payment credentials, and account information, making privacy-enhancing prompting techniques ineffective. Another concern is the high frequency of data access and potential leakage. A user searching for medical devices may visit a few websites, but an automated agent could query dozens, embedding the user's medical interests into multiple tracking systems. If these queries reach malicious sites, sensitive health information could be exposed. Similarly, an agent checking flight prices could unknowingly share location data across multiple services, increasing surveillance risks. These automated interactions contribute to detailed behavioral profiles, which, if leaked or misused, could lead to data exploitation and unauthorized inferences about personal habits. *

Diminished Privacy and Security Control: GUI agents improve efficiency but reduce user control, making privacy risks harder to assess. Unlike direct interactions where users can adjust their inputs in real time, GUI agents operate autonomously, requiring users to trust their decision-making before the interaction begins. For instance, automating tax filing requires providing credentials, financial documents, and personal data. While executing tasks, users may remain unaware of how their data is processed, stored, or shared. Similarly, an agent recovering social media accounts might input security answers or recovery codes without user oversight, increasing unauthorized access risks. This lack of visibility makes it difficult for users to assess and mitigate privacy threats, leading to potential misuse.

Insufficient Guardrails: Privacy safeguards are often overlooked in GUI agents, making them vulnerable to adversarial attacks. For example, Claude's Computer Use agent unknowingly shared a fake driver's license number with a phishing website. Following user instructions to obtain a discount, the agent failed to recognize the fraudulent site or question the unusual request. GUI agents processing structured files, such as HTML or APK, are also at risk. Environmental Injection Attacks (EIA) exploit these weaknesses by injecting malicious content that dynamically adapts to the agent's environment. In one study, an agent was tricked into entering personal data into an invisible field containing malicious instructions, unknowingly leaking sensitive information. Such vulnerabilities highlight the need for integrated privacy safeguards to prevent data breaches.

Challenges in Evaluations: Despite privacy concerns, GUI agent evaluations primarily focus on performance, assessing effectiveness (e.g., task completion rates) and efficiency (e.g., speed and resource use). While some studies measure safety, they often address immediate risks rather than nuanced privacy concerns. Research has shown that models with lower leakage rates often perform worse in helpfulness, suggesting a tradeoff where some agents prioritize responsiveness over privacy. A major challenge in assessing privacy risks for GUI agents is their strong dependence on context. Privacy calculus theory suggests that users weigh risks and benefits based on perceived rewards and system trust. Contextual integrity theory highlights that privacy decisions depend on specific circumstances, including information type, user-system relationships, and task context. This variability complicates standardized risk assessments, requiring a context-aware approach that considers individual privacy value judgments. To address this issue, we advocate for a human-centered evaluation framework for trustworthy GUI agents. Unlike traditional automation, GUI agents dynamically interpret and interact with interfaces, necessitating privacy safeguards alongside performance optimization. We propose three key actions: (1) human-centered evaluation for privacy risk assessment, (2) integrating privacy measures into agent development, and (3) enhancing user awareness through in-context consent mechanisms.

Call for Action: To ensure trustworthy GUI agents, we recommend the following:

Human-Centered Evaluation for Privacy Risk Assessment: Unlike traditional automation, GUI agents require in-context evaluations with user oversight. Increasing system complexity and invisible data transmissions necessitate systematic privacy risk assessments across UI perception, intent generation, and action execution. Since users may lack expertise in identifying privacy risks, evaluation frameworks should enhance their ability to recognize and manage risks. The multimodal nature of GUI agent outputs also increases cognitive burden, complicating oversight. Evaluations should focus on unintended data exposure, transparency, and minimizing oversight challenges to ensure safe and trustworthy automation.

Enhancing Users' Privacy Awareness with In-Context Consent: GUI agents should include explicit warnings and in-context consent mechanisms. Since users may struggle to understand privacy risks and tend to overtrust AI, agents should retrieve and process privacy policies, offering contextualized explanations and actionable guidance. Structured consent requests should precede privacy-sensitive actions—such as sending emails or conducting transactions—ensuring user control. Configurable privacy settings should allow users to balance automation convenience with data protection based on their needs.

Integrating Privacy Measures into Agent Development: Privacy safeguards must be embedded in prompt-based and training-based GUI agent development. In prompt-based methods, data protection should be enforced through explicit instructions, limited data retention, and required user consent before accessing sensitive information. In training-based methods, privacy protections should be integrated throughout development: pre-training with privacy-focused datasets, fine-tuning to prevent breaches, and reinforcement learning to reward protective behaviors while penalizing unauthorized data exposure. These measures ensure privacy is a core design principle, fostering informed oversight rather than blind trust.

Conclusion: The increasing adoption of GUI agents presents both opportunities and risks. While they enhance automation and accessibility, their ability to process and interact with sensitive data raises significant privacy concerns. Existing evaluations primarily focus on performance, often overlooking nuanced privacy risks. We call for a human-centered approach to evaluating and developing GUI agents, ensuring privacy safeguards are integrated into their design and deployment. By implementing structured risk assessments, enhancing user awareness, and embedding privacy measures in development, we can build GUI agents that are not only efficient but also trustworthy and secure.

#112: In Attaining the Unattainable: Agency and Creativity in AI-generated Liminal Space

Authors: Xinyue Liu

Abstract: The rise of Generative AI (GenAI) begets a novel modality of human-machine interaction featuring enhanced feedback, reduced human control and intensified machine agency. It also introduces affordances and constraints to human creativity, as well as opportunities and challenges to practitioners in the creative industry. Artists and technicians experiment with the emerging GenAI to generate content, which usually follow the pipeline of 1) human-intended prompt; 2) machine-generated content; 3) human-finalized outcome. Noticeably, there are inevitable gaps between the expected result with human intention and the generated content by AI models. The process of finalizing the outcome, in particular that of patching the gap in-between, involves not only the creative agency of humans, but also the inspiration that humans draw from the machine-generated contents. The process of attaining "the unattainable" (Bellour 1975 & 2010) therefore reflects the immaterialized machine agency that takes place in a cognitive space.

Drawing from ethnographic research inside an AI-powered tech start-up, along with a close study of an art-practice of a digital artist with the aid of Generative AI, this paper initiates a semiotic reading of the AI-involved creative process. This paper advocates for the necessity of theorizing the gap between the human expectation and machine-generated content, therefore facilitating the exploration of the potential creativity and the dynamic of human/machine agency that are conceived therewith. Building upon Victor Tuner’s “liminality” (1969), an anthropological concept that signifies the concurrent ambiguity and possibilities in the transitional phase, this paper explores the potential of reading the gap as an AI-generated liminal space. By integrating into the broader discussion of human-machine co-creation, this paper investigates possibilities of a new modality of creativity in the age of AI and advocates for ethical application of AI in the creative industry with greater fairness, transparency, and accountability.

#121: Ethically And Efficiently Leveraging AI To Effectuate Employment Hiring Decisions

Authors: Alex Karasik

Abstract: Recruiting and retaining talent is the lifeblood of employers. While traditional hiring practices such as reviewing resumes and interviewing candidates remain building blocks of the recruitment process, in recent years, the artificial intelligence wave crested upon the employment decision-making landscape. Businesses are turning to AI to streamline otherwise burdensome tasks such as locating candidates and filtering resumes. AI is also being implemented into the interview process, helping companies evaluate candidates by assessing their behavioral cues, personality traits, and accuracy of responses. Inherit in the efficiencies of AI use in the employment decision-making process is the imminent risk of bias. While the AI can read resumes faster than humans and quantify behavioral traits that a human might overlook, this emerging technology is still grappling with how to eradicate bias. For instance, how does an AI interview rank a prospective candidate with a speech impediment without knowing that candidate may have a disability? Or what if a resume screening tool filters out a disproportionate number of applicants from a particular demographic?

While employers and AI technology developers grapple with these scenarios, some businesses are already in the crosshairs of AI-related employment discrimination litigation. In 2023, the United States Equal Employment Opportunity Commission (the “EEOC”) reached its first ever AI-based anti-discrimination settlement for $360,000, to resolve allegations that a tutoring company discriminated against applicants on the basis of age by using automated systems in recruiting software that filtered out older candidates. Similarly, a discrimination lawsuit currently pending in a federal court in California, Mobley v. Workday, involves allegations that Workday’s AI-powered applicant screening tools

discriminate on the basis of age, race, and disability. These cases are examples of alleged AI-related bias rapidly percolating into high-stakes litigation.

Problem Statement: How can businesses ethically and efficiently use AI in employment decision-making processes without subjecting applicants or employees to bias or discrimination?

Translation Evidence: Employment discrimination is an issue that impacts nearly every industry and geographic region in the United States. In fiscal year 2024, the EEOC reported that 88,531 new discrimination charges were filed, an 9.2% increase from the prior year. According to the ACLU, 70% of companies – including 99% of Fortune 500 companies – are using AI-based and tools in their hiring processes, particularly to fill low-wage positions. Taken together, these statistics suggest that the increased use of AI in the hiring and employment decision-making processes may be adversely impacting workers who are members of protected classes.

RISE Dimensions: To solve the above Problem Statement, employers need to take a proactive and hands-on approach to incorporating AI tools into the workplace. Through my experience counseling businesses of all sizes on how to responsibly incorporate AI into their workplaces, the following five strategies emerged as the best tools to fracture the risk of bias and discrimination.

1. AI Committee. The first step for any organization wishing to implement AI-tool into HR functions is to establish a committee of stakeholders with diverse perspectives. This can include upper management, HR, Legal, CIO/IT, leaders in different departments, and others with unique perspectives. The AI Committee can develop a robust policy that serves that ethically serves the business needs of the organization.
2. AI Policy. Crafting a clear and cohesive AI policy sets the foundation to create a level playing field for AI use in employment processes. This also protects the organization in other areas such as data privacy and compliance.
3. Employee/Manager Training. Both workers and decision-makers, particularly those involved in HR functions, need to be trained on how to equitably implement the AI Policy, and understand its interplay with other policies such as EEO. For instance, if an applicant asks for an accommodation during an interview tool that uses AI, training the interviewer on how to accommodate that request is crucial.
4. Vendor Communications. Prior to using an AI tool in the employment decision-making process, it is imperative that businesses communicate with vendors to understand how AI models were trained. Businesses need to ask pointed questions, for instance, "How does this language model evaluate foreign accents?," and "Have these tools been vetted such that hypothetically disabled applicants have been able secure jobs during training tests?"
5. Audits. As the EEOC's Artificial Intelligence and Algorithmic Fairness Initiative consistently suggests, employers should audit AI tools to determine whether their output has a disparate impact on certain demographics of applicants. Audits should be done regularly, both before and after the implementation of AI tools. If an audit determines that an AI tool has an adverse impact, the employer should address the issue with the AI vendor. Data is paramount in terms of diagnosing potential large-scale bias, and in turn preventing class action litigation. Audits identify this data.

Conclusion: Following the "storm" of potential AI-generated employment decision-making bias is a "rainbow" of optimism: AI has great potential to reduce workplace bias, such as when an interviewer may have pre-conceived notions about a certain demographic. But similar to a counter-terrorist agent who prevents an attack or a referee in sports who calls a game fairly, the positive impacts of AI in employment practices are difficult to quantify and rarely celebrated. If personnel and AI tools are trained properly on how to remove bias from employment decision-making processes, AI tools could reduce employment barriers. The businesses who choose to ethically adopt AI into employment practices will ultimately bear the fruits of AI's efficiency, creating a qualified and inclusive workforce that mirrors our potent and innovative society.

Track: Reimagining Global Governance and Policymaking

#2: A Model of Algorithmic Wage Determination

Authors: Abhishek Ray

Abstract: Should AI-powered employee monitoring systems let employers set monitoring levels? We explore this question in light of the fundamental AI alignment dilemma as firms increasingly use AI-powered 'bossware' technology. AI-powered bossware technologies, popular with various employers, collect and process workers' efforts, compromising workplace privacy and mental health. While lawmakers have sought to limit bossware through regulation, employers have consolidated data collection through intrusive ways. Thus, setting monitoring levels is a contentious issue. We address this problem by proposing a novel principal-agent relational contract framework that consists of a social planner (lawmaker), principal (employer), and agents (employees) that, under monitoring, may be tech-loving (thriving) or tech-averse (suffering). We use this setting in a benchmark model (without monitoring technology) and a comprehensive model with employer- or social planner-implemented monitoring levels. The equilibrium analysis yields several insights.

The benchmark model reveals that only one incentive contract is viable when an employer trusts employees, where tech-loving employees are paid less than tech-averse individuals. However, when an employer sets monitoring levels, tech-averse employees are paid more than tech-lovers to encourage desired efforts to maximize profitability or productivity. Still, tech-loving employees earn the same whether they increase productivity or profitability, suggesting businesses should prioritize tech-averse employee incentives. In contrast, the comprehensive model indicates that social planners setting monitoring levels considerably affects contractual results. If technology setup costs are cheap, the employer offers agents the same contracts, whether the social planner is utilitarian or Rawlsian. Low setup costs let social planners choose utilitarian or Rawlsian monitoring levels; the principal is indifferent and proposes the same contracts. However, when setup costs are significant, utilitarian monitoring levels may not be lower than Rawlsian. Tech-lovers are always compensated more than tech-averse under Rawlsian or Utilitarian. Unlike the benchmark model, the comprehensive model recommends letting the social planner set monitoring levels that favor tech-loving over tech-averse employees. Our research thus adds to the fairness and privacy aspects of AI use in businesses.

#25: Distilling heterogeneous treatment effects: Stable subgroup estimation in causal inference

Authors: Tiffany Tang, Melody Huang and Ana Kenney

Abstract: Recent methodological developments have introduced new black-box approaches to better estimate heterogeneous treatment effects; however, these methods fall short of providing interpretable characterizations of the underlying individuals who may be most at risk or benefit most from receiving the treatment, thereby limiting their practical utility. In this work, we introduce causal distillation trees (CDT) to estimate interpretable subgroups. CDT allows researchers to fit any machine learning model to estimate the individual-level treatment effect, and then leverages a simple, second-stage tree-based model to distill the estimated treatment effect into meaningful subgroups. As a result, CDT inherits the improvements in predictive performance from black-box machine learning models while preserving the interpretability of a simple decision tree. We derive theoretical guarantees for the consistency of the estimated subgroups using CDT, and introduce stability-driven diagnostics for researchers to evaluate the quality of the estimated subgroups. We illustrate our proposed method on a randomized controlled trial of antiretroviral treatment for HIV from the AIDS Clinical Trials Group Study 175 and show that CDT out-performs state-of-the-art approaches in constructing stable, clinically relevant subgroups.

#35: Can an Impersonal and Objective Algorithm Become a Better Law Enforcer? Insights from Confucian Criticism of Chinese Legal Philosophy and Practice

Authors: Liang Cai

Abstract: In this transformative age, AI has become a designer of regulations and an enforcer of laws and policies. The corporate world increasingly embraces AI to monitor, reward, and discipline employees, valuing its objectivity, impartiality, and efficiency. Humankind may finally be satisfied with a creation immune to human biases and emotions, as an orderly world—operated with precision by algorithms—emerges on the horizon. Visualizing a world of order untouched by human interference—free from subjective biases, conflicting values, and complex emotions—was also a dream of Chinese legal philosophers. Shang Yang and Han Feizi eloquently articulated the effectiveness of performance-based laws and regulations. By strictly evaluating performance against legal objectives without granting law enforcers any discretionary power, they believed that no one would dare deceive the ruler or violate the law, thereby creating a crime-free utopia. The world envisioned by Chinese Legalist philosophers closely resembles one governed by algorithms. By examining both archaeologically excavated legal statutes and cases alongside transmitted sources, I demonstrate that when brutal instrumentalism and idealism were applied to the real world, they produced a monstrous system that distorted justice. A system that neither tolerates human error nor allows discretion in applying laws and regulations may achieve high efficiency but inevitably punishes large numbers of people, including those diligently devoted to their work. When punishment is neither deserved nor just, resentment toward the law arises, and sympathy grows for the condemned. This study presents a historical case to provoke reflection on the dangers of perfectionism—an ideal that AI may one day pursue—when applied to the complexities of the real world.

Governance should aim to serve human beings rather than treat them as mere tools to maximize efficiency within a rigid and perfect order. If relying on AI for governance due to its efficiency is unavoidable, we must identify strategies to mitigate the risks of instrumentalizing human beings. In designing AI governance, we should incorporate tolerance—allowing for permissible limits of error or deviation. This concept, crucial in fields such as engineering, manufacturing, and quality control, helps maintain consistency and reliability. Additionally, it's important to establish higher principles, such as prioritizing human life and compassion, to guide the regulations within AI systems. Furthermore, it is essential to create an appealing system with AI governance that allows educated and virtuous individuals to exercise their discretionary power. Finally, we should consider how to infuse virtue into AI training, aiming to develop a virtuous AI system much like we strive to cultivate virtuous human beings.

#36: Building AI Text Classifiers with Peacebuilders: A Human-AI Collaboration to Improve Conflict Analysis and Resolution

Authors: Julie Hawke and Will O'Brien

This article investigates the outcomes of integrating peacebuilders into the collaborative development of AI-driven text classification models for conflict analysis and resolution. Drawing from participatory action research conducted through the Automatic Classifiers for Peace (ACfP) initiative, we analyze how involving peacebuilding practitioners directly in the iterative design, annotation, and evaluation phases of AI-drive text classification model development impacts their awareness, attitudes, and willingness to adopt AI technologies in their professional practices. Additionally, the study evaluates improvements in the quality and contextual relevance of classification models created through this human-AI collaboration. Conducted over two four-month cycles, this ongoing research aims to provide empirical evidence supporting the efficacy of participatory AI development methods, demonstrating enhanced classification performance and greater practitioner ownership and trust in AI-generated insights.

#38: The Perks and Perils of Machine Learning in Business and Economic Research

Authors: Tom Dudda and Lars Hornuf

Abstract: Because of its non-linear and complex nature, machine learning (ML) offers significant theoretical advantages over traditional statistical models in addressing prediction problems in business and economics. It is therefore not surprising that researchers increasingly apply ML models to address research questions in business and economics (Figure 1). However, many problems are in reality rather simple and sometimes linear in nature. Using ML models to approach these problems can be inefficient given the higher costs associated with ML models in terms of time and energy consumption, adverse environmental impact, and lack of explainability of results compared to traditional regression models. Given the rapidly rising number of ML-related publications in business and economic research in recent years, we hypothesize that published articles in these disciplines increasingly employ ML models for predictive research problems that often do not have the content complexity that requires researchers to rely on more resource-intensive models. In such cases, these models will frequently perform only marginally better than conventional, less time- and energy-consuming methods that have been traditionally used in the literature. Researchers might refrain from reporting the predictive performance of less complex conventional methods if they yield similar results to ML models, given that marginal improvements might not justify the higher financial and environmental costs of using more complex models.

Using novel ML methods might, in contrast, attract more attention, leading to higher chances of publication and a higher number of expected citations. We examine predictive ML studies from the 50 highest-quality business and economic journals published between 2010 and 2023. First, we investigate their transparency regarding the predictive performance of ML models compared to less complex traditional statistical models, such as linear or logit regression, that require fewer resources in terms of time and energy. Second, provided that the studies report comparable results for both ML and established traditional statistical models, we compare the reported predictive performance of both model types. Finally, we examine whether the transparency about and the extent of the relative performance improvement through ML models is associated with the impact that an article generates, as measured by its citation count.

Results: Out of 56,262 articles published between 2010 and 2023 in journals of the FT50, we manually identified 1,211 articles that involve ML. The sample for our main analysis consists of 203 studies that apply at least one ML model to predict a variable that is of central interest for answering one of the main research questions of the article. We find that 28% of articles do not benchmark the predictive performance of the employed ML models against traditional statistical models.

This finding is only partly explainable through the use of innovative data sets (like text and images), which are potentially less suitable for traditional prediction models. Neglecting to report comparable results for traditional statistical models makes it difficult to assess the true economic value of using a more complex and resource-intensive ML model. We encounter substantial differences between research disciplines but find no evidence that the seniority of the authors or the size of the author team affects transparency about the model performance. Second, studies reporting results for traditional statistical models predominantly state a strong outperformance of the best-performing ML model over the best-performing traditional benchmark, pointing towards a potential publication bias and selection of weak benchmarks. The outperformance is, on average, reduced by 65%, often even turning negative, when we compare the average performance of all reported ML models in a study against the best traditional model.

Lastly, we find that the transparency of published articles about the relative performance of ML models compared to traditional statistical models is positively related to their citation count. Studies that report results for traditional benchmarks receive, on average, 2.7 to 6.8 more citations per year than studies that do not report such benchmark results. This effect is sizeable considering the 8.1 citations per year that the average article published in the journals of our sample garnered between 2018 and 2023. We argue that studies transparently reporting benchmark results are methodically more rigorous, which might also be an indicator of the general quality of the study, which ultimately results in more citations.

Alignment with RISE dimensions: Our study highlights important implications for future research based on resource-intensive AI models, also outside of business and economics. Our results indicate that, for the average ML model, there is often relatively little gain in predictive performance over traditional benchmark models. Time and energy seem to be required to achieve considerably improved predictions with well-trained ML models. However, we contend that, due to opaque reporting practices, it often remains unclear whether the predictive gains justify the increased costs of more complex, resource hungry models. We advocate for standardized, transparent model reporting that relates predictive gains to the efficiency of ML models compared to less-costly traditional statistical models. We consider our study to be particularly aligned with the RISE dimensions Responsibility and Ethics.

Overall, our research discusses responsible research with AI models with regard to resource consumption and environmental impact, and emphasizes the importance of ethical AI research practices in terms of a transparent reporting of relative model performance and model costs compared to less-complex and less-costly but potentially still powerful alternatives.

#52: Assessing Science Use in AI Policymaking: A Case Study of California

Authors: Frederick Boehmke, Bruce Desmarais, Jeffrey Harden and Sarah Rajtmajer

Artificial Intelligence (AI) policy is a rapidly emerging field characterized by complexity and rapid technological evolution, presenting substantial challenges to policymakers unfamiliar with the detailed scientific knowledge underpinning these innovations. We study, and develop tools to automatically assess, how scientific evidence informs state-level AI policy decisions. By zeroing in on California—a state that, as of the end of 2024, has already passed nearly 100 bills related to artificial intelligence—this project captures the role of science in policy formulation, adaptation, and diffusion within a highly active and influential state. Policy diffusion, or how policy ideas spread among lawmakers and jurisdictions, serves as a central theme of this initiative.

Given AI's novelty and its associated complexities, state legislators frequently rely on external scientific expertise and peer interactions to navigate uncertainties. California provides a compelling case due to its early adoption of AI-related legislative proposals and historical position as an innovative leader in policymaking, offering unique insights into the micro-level mechanisms through which policymakers engage with scientific information and implement new technologies into law. We integrate natural language processing (NLP) methods to explore how scientific evidence is referenced, communicated, and utilized by legislators. It systematically identifies and evaluates the quality and applicability of scientific citations appearing in legislative texts, committee hearings, and public statements. Using advanced NLP tools, including large language models (LLMs), we assess whether policy-relevant scientific claims cited by legislators align with broader scientific consensus or reflect disputed or selective use of evidence.

By applying advanced computational techniques, including the automated extraction and assessment of scientific claims cited by legislators, the study advances the state of the art in evaluating the quality and impact of science-policy interactions. This methodologically innovative approach facilitates the accurate identification of scientifically robust claims versus selectively cited or weaker evidence, thereby clarifying the extent to which credible science influences policymaking. The outcomes of this project carry significant implications beyond academic theory, offering actionable insights for policymakers, researchers, and advocacy groups committed to evidence-based governance. The methods help us to understand effective strategies for promoting scientifically informed AI policy, enhancing the broader societal impact of policy measures designed to manage the ethical, economic, and societal implications of AI.

#117: The Impact of Artificial Intelligence on Labor Markets: A Forecasting Model for Significant Economic Events

Authors: Andrew Bond and Cadence Rand

Abstract: This study introduces a scalable, regionally adaptive framework for forecasting the impact of artificial intelligence (AI) on labor markets, designed to inform policy responses at local, national, and global levels. Recognizing the limitations of deterministic models and unstructured speculation, an ensemble of advanced large language models (LLMs) is used to classify over 1,000 occupations from the Standard Occupational Classification (SOC) system using structured prompts and task-level data from O*NET. Rather than estimating continuous levels of AI disruption, our method applies quantization—discretizing occupations into five categories: High Displacement, Displacement, Neutral, Growth, and High Growth. This approach leverages the deep reasoning capabilities of LLMs to consistently evaluate job characteristics such as task automation potential, social complexity, and creative requirements. Cross-validation across multiple frontier LLMs ensures robustness and minimizes individual model bias. These classifications serve as inputs to a stochastic Monte Carlo simulation that models employment outcomes over a five-year period under varying AI adoption scenarios. Simulations incorporate regional employment trends and assign randomized impact factors within each job category to generate a distribution of plausible labor market trajectories. Stochastic modeling was used as it better captures the uncertainty inherent in real-world labor markets, producing a more robust range of possible outcomes. Preliminary results reveal significant geographic variation in occupational vulnerability to AI across U.S. regions, underscoring the need for flexible and proactive policy design. Targeted interventions, particularly in regions with high concentrations of automatable jobs, will be essential to mitigating displacement risk while leveraging AI's economic potential. This research offers a repeatable framework for labor market forecasting that is more informative, enabling decision-makers to prepare for and adapt to the evolving dynamics of AI-driven transformation.

#128: Stochastic Justice: Inherent Non-Deterministic AI Inference in Legal Decision-Making

Authors: Or Cohen-Sasson, Irit Opher and Or Cohen-Sasson

Problem Statement: Legal systems worldwide operate on a fundamental principle: similar cases should yield similar outcomes. This consistency is essential for legitimacy, fairness, and predictability in legal processes. However, modern AI systems, particularly large language models (LLMs), exhibit inherent non-determinism—producing different outputs for identical inputs across iterations. This creates a fundamental tension with legal consistency requirements. Our research addresses a critical challenge at the intersection of law and technology: determining the upper bounds of legal automation given the non-deterministic nature of current generative AI technologies, and understanding the implications of AI-induced inconsistency in legal decision-making contexts.

AI Innovation: Our methodological innovation is threefold: First, we develop a systematic categorization framework for different types of legal consistency in AI inference, establishing a taxonomy that bridges technical and legal domains. Second, we perform technical analysis of non-determinism sources in AI systems, identifying how frameworks like Mixture of Experts (MoE), non-linear computational methods, inherent randomness, and hardware complexities contribute to this phenomenon. Third, we employ novel testing methodologies to quantify AI inconsistency rates across different legal tasks and contexts, providing empirical evidence of the scope and nature of the problem.

Translational Evidence: Our findings have significant deployment and policy implications. We establish evidence-based boundaries for appropriate AI implementation in legal contexts, identifying which legal functions are suitable for automation and which remain vulnerable to problematic inconsistency. Our research provides decision-makers with a framework to assess existing and proposed AI legal tools, potentially informing regulatory approaches and technical standards for AI systems in legal settings. For legal technology developers, our work offers guidance on necessary consistency thresholds and evaluation methodologies for AI legal tools, ultimately ensuring more reliable and trustworthy legal technology deployment.

RISE Dimensions: This research primarily addresses Responsibility and Ethics in AI development. By identifying and quantifying AI-induced inconsistency in legal contexts, we promote responsible development of legal AI tools that align with fundamental principles of justice. From an ethical perspective, our work examines fairness implications when similar legal questions receive different AI-generated answers, addressing core concerns of procedural and substantive justice. Additionally, our research contributes to Safety by identifying potential harms from inconsistent legal decision-making, particularly for vulnerable populations who may lack resources to challenge AI-influenced decisions. By examining how non-determinism affects different types of legal questions, we also highlight Inclusion concerns regarding equitable access to consistent legal outcomes across demographic groups and case types.

Research Abstract – Further Information This research investigates a fundamental tension between artificial intelligence technology and legal consistency principles. Legal systems worldwide are built on the premise that similar cases should yield similar outcomes—a principle embodied in doctrines like stare decisis that ensures legitimacy, fairness, and predictability. However, our empirical testing reveals that current AI technologies, particularly large language models (LLMs), exhibit inherent non-determinism that produces different outputs for identical inputs across iterations. This technological characteristic creates a concerning incompatibility with legal consistency requirements that has significant implications for AI adoption in legal contexts.

Our study systematically examines this tension through empirical testing of multiple legal scenarios, demonstrating that output variations from AI models are often inconsistent in legally meaningful ways. The non-determinism stems from several technical factors, including different frameworks such as Mixture of Experts (MoE), non-linear computational methods, inherent randomness, hardware-related complexities, and ongoing model updates. Our research quantifies inconsistency rates across various legal tasks and model types, providing concrete evidence of the scope and implications of this phenomenon.

The paper makes four primary contributions to understanding this critical intersection of technology and law: (1) systematic categorization of different types of legal consistency in AI inference; (2) technical analysis of the sources of non-determinism in current AI systems; (3) testing the extent of AI non-determinism and inconsistency in legal decision-making across different models; and (4) examining fundamental similarities and differences between human- and AI-induced non-determinism and inconsistency. Our investigation specifically addresses key questions about AI implementation in legal contexts, including: What is the legal inconsistency rate in AI inference across different tasks? What are the effects of such AI inconsistencies in different legal contexts? What are the effects, if any, of human-induced inconsistency in legal decision-making on the boundaries of legal automation? By analyzing these questions through both technical and legal lenses, we establish evidence-based boundaries for appropriate AI implementation in legal contexts and provide decision-makers with frameworks to assess existing and proposed AI legal tools. This research is particularly timely as legal institutions worldwide increasingly explore AI adoption to improve efficiency while maintaining fundamental principles of justice and consistency.

#138: Unpacking the "Good" in UN AI for Good Projects Using Computational Grounded Theory

Authors: Jose Marichal

Abstract: This work is a empirical inquiry into the presentative given to the The United Nations' AI for Good initiative. This project brings together hundreds of projects and presentations under the auspices of the International Telecommunications Unit (ITU), the UN’s technology agency. The goal is to link AI innovations to the 17 Sustainable Development Goals (SDGs). Scholars who have studied the work of the initiative recognize that the term "good" is often left intentionally vague in order to accommodate a large number of projects. Scholars have raised concerns that too vague of a definition of “the good” runs the risk of treating political/social problems as purely technical (Li 2007, Morozov 2013), to rationalize “suffering reduction” without addressing root causes (Madianou, 2021), and to treat global south communities as solely investigatory sites (Haelewaters et. al. 2021). Despite these criticisms, there exists little empirical work that seeks to map the discourse structure of various AI good projects. What are the different ways in which the good is defined? Who is more likely to have a comprehensive, non-technosolutionist view of the good?

Our research seeks to systematically examine the differences in how presenters in the AI for Good initiative frame "the good" by employing a mixed-method approach called computational grounded theory (Nelson 2017). Computational

Grounded Theory (CGT) has three components. The first is using traditional AI tools, in our case BERTopic topic modeling, to identify patterns in a corpus of data. The second is doing a close reading of data to elaborate on the patterns found by the AI. Finally, the new concepts are used to train a classifier to laber the entire corpus of data. We report on the first and second parts of this project. We ran a BERTopic model on a corpus of 417 video transcripts downloaded from the AI for Good YouTube channel. We used BERTopic to identify key themes, which included areas aligned with specific SDGs, technical areas, and meta-themes about AI itself. We will then conduct a close reading/watching on representative videos from the topic model.

Preliminary work in this regard has revealed two dominant justification frames: info-solutionism and auto-solutionism. Info-solutionism assumes that more data will inherently lead to better outcomes, while auto-solutionism frames inefficiency as the core problem, solvable via automation. These frames have significant implications for the impact of AI on society, particularly in terms of equity, bias, and labor displacement. We propose to refine and expand on these categories and present this work at the RISE conference. This work is aligned with the goals of the RISE conference. Through this work, we hope to Identify patterns and trends in the data that can inform the development of more effective and equitable AI solutions. Very few studies have systematically examined the justifications being used to legitimate AI solutions. Helping UN and ITU agents better identify the set of arguments presented will lead to better questions and a more critical and rigorous process of vetting the appropriateness and limitations of AI solutions for specific policy domains. We believe this will ultimately inform the development of more inclusive and participatory AI solutions that prioritize human flourishing, with a focus on involving communities in the design and development process.

This work is consistent with the aims of the RISE conference because it centers prioritizing human well-being. In our preliminary work, we've identified justifications of AI projects that appear to put SDG's in conflict. For example, by emphasizing auto-solutionism in wind energy robotics and agricultural robotics, researchers run the risk of ignoring the job losses and community devastation that result from widespread adoption of AI solutions. By highlighting the importance of involving communities in defining "flourishing" and ensuring that AI development prioritizes human well-being, we hope to contribute to a more nuanced and equitable understanding of the role of AI in achieving the SDGs. Our extended analysis aims to contribute to a more nuanced understanding of the "good" in AI for Good projects, highlighting the trade-offs and potential harms associated with different justification frames. By exploring the implications of our research for the future of AI development, we hope to inform the development of more inclusive and participatory AI solutions that prioritize human flourishing. We believe that our research has the potential to make a significant contribution to the field of AI development, and we look forward to interacting with scholars and practitioners at the RISE conference that can help us improve and extend our analysis.

#174: AI-Driven Detection of Fake News through Emotional Representation Analysis

Authors: Vitali Herrera Semenets, Lázaro Bustio Martínez and Jan van den Berg

Problem Statement: The proliferation of fake news poses a critical societal challenge, undermining trust in media, influencing public opinion, and threatening democratic processes. Traditional detection methods often rely on linguistic features (e.g., TF-IDF), but these may overlook the emotional manipulation tactics central to fake news dissemination. This work addresses the need for innovative, interdisciplinary approaches to improve fake news detection while balancing ethical and operational trade-offs (e.g., recall vs. precision).

AI Innovation: We propose a novel representation method for fake news detection based on emotional analysis, leveraging machine learning (ML) models to capture the affective dimensions of text. Compared to TF-IDF baselines, our approach achieves exceptional recall (up to 1.000 with SVM), ensuring minimal false negatives critical for high-stakes scenarios where fake news must be flagged reliably. The innovation lies in prioritizing emotional features (e.g., arousal, valence) as proxies for manipulative intent, complementing traditional lexical methods.

Translational Evidence: Experiments show the emotional representation excels in recall (99–100% for Logistic Regression/SVM), outperforming TF-IDF in detecting fake news, albeit with lower precision. This suggests immediate applicability in settings where false negatives are costlier than false positives (e.g., public health misinformation monitoring). Future integration with persuasion principles and sentiment analysis could further enhance robustness.

The solution is scalable for deployment in social media moderation tools, news aggregators, or policy frameworks aiming to curb misinformation.

RISE Dimensions: Our approach aligns with three key dimensions of the RISE framework: Responsibility, Ethics, and Inclusion. By prioritizing recall, the model reduces societal harm from undetected fake news, adhering to the principle of Responsibility in risk mitigation. From an Ethical standpoint, we transparently address the precision-recall trade-off, advocating for context-aware deployment (e.g., maximizing detection in scenarios where misinformation threatens public safety). Finally, emotional analysis introduces an Inclusive component, as affective patterns are more universal than linguistic ones, potentially reducing cultural or language biases inherent in traditional text-based methods. This threefold alignment reinforces the socially conscious design of our solution.

Conclusion: This work bridges AI, psychology, and media studies to tackle misinformation, demonstrating how emotion-aware ML models can complement existing tools. Future directions will explore context-aware deployment strategies, such as dynamic threshold adjustment based on content risk levels (e.g., stricter filtering for political misinformation). By combining emotional, linguistic, and persuasive features, this expanded model could set a new standard for explainable, ethical, and adaptive misinformation detection aligning with RISE principles while addressing real-world policy and moderation needs.

Track: AI for Safety and Safety of AI

#37: On Implicit Social Biases in LLMs' Chain-of-Thought Reasoning

Authors: Deng Pan, Joe Germino, Yihong Ma, Elizabeth Daly, Nuno Moniz and Nitesh Chawla

Abstract: Large Language Models (LLMs) employing Chain-of-Thought (CoT) reasoning have demonstrated significant improvements in complex problem-solving and decision-making. However, the fairness implications of CoT remain underexplored, raising concerns about potential biases that emerge in multi-step reasoning processes. This study examines how implicit social biases propagate through CoT reasoning, leading to disparities in outputs based on implicitly inferred demographic or contextual factors. We introduce a fairness evaluator to quantify fairness in CoT reasoning steps and propose the critique-and-reflection mitigation strategy to reduce bias in the reasoning processes.

Introduction: Large Language Models (LLMs) using Chain-of-Thought (CoT) reasoning have significantly enhanced AI's ability to handle complex reasoning tasks. Despite the advances, an important yet underexplored concern is the propagation of implicit and explicit social biases through intermediate CoT steps. Such biases can amplify through multiple reasoning steps, resulting in outcomes that unfairly favor or disadvantage particular socio-demographic groups. This research investigates bias propagation in CoT reasoning. We propose fairness metrics designed to quantify and analyze biases at intermediate reasoning stages, providing transparency into how these biases evolve. Furthermore, we adopt the critique-and-reflection mitigation strategy, enabling iterative bias correction and promoting fairness in LLM reasoning processes.

Methodology: Proposed Method We adopt the critique-and-reflection framework for fairness detection and mitigation in multi-step CoT reasoning. Our framework involves two core components: Fairness Critique Module: A fairness evaluator that systematically quantifies bias at intermediate reasoning steps by counterfactual analyses, entailment scores, or sensitive attribute prediction proxies. This approach allows explicit measurement and tracing of bias propagation through multi-step reasoning. Reflection-based Bias Mitigation: An iterative self-reflection mechanism whereby LLM-generated CoT steps are evaluated for fairness using the fairness critique results. When biases are detected, the model leverages these reflections to refine or reject previous reasoning chains. This reflective process, inspired by recent advancements in mitigating hallucinations, continuously integrates fairness feedback into subsequent reasoning iterations until equitable and unbiased outcomes are achieved. Experimental Design We validate our proposed framework using two types of datasets: (1) traditional tabular datasets with clearly defined sensitive attributes, which we transform into natural language question-answering datasets using LLMs; and (2) specialized datasets designed explicitly for social bias detection, such as the BBQ dataset. Our experimental analysis includes: Examination of implicit bias propagation patterns across multi-step CoT reasoning processes. Evaluation of fairness detection accuracy using different critique methodologies, including counterfactual analyses, entailment scoring, and sensitive attribute proxies. Assessment of the effectiveness and robustness of our reflection-based mitigation strategy, specifically evaluating bias reduction capabilities and the preservation of overall reasoning quality.

Expected Outcomes: We expect to demonstrate that implicit biases are not only propagated but amplified during CoT reasoning. Our fairness critique module will effectively identify biases at intermediate reasoning steps, offering transparency into how biases develop over the course of reasoning. Furthermore, we anticipate the reflection-based mitigation mechanism to significantly reduce implicit bias, enabling the generation of more equitable reasoning outcomes without compromising model performance and interpretability.

Broader Impact: Our critique-and-reflection framework has broad implications for the responsible deployment of AI across various critical decision-making domains such as healthcare, employment, legal systems, and education. By explicitly addressing and mitigating biases in reasoning, this approach can significantly reduce algorithmic discrimination, particularly benefiting marginalized and underrepresented communities. Additionally, the iterative and transparent nature of our framework empowers policymakers, researchers, and practitioners to more effectively audit and validate AI fairness, fostering greater public trust and supporting ethical AI practices.

#39: Data-centric AI for the Early Detection of Suicide Risk in Chilean Schools

Authors: Paul Escapil-Inchauspé, Gonzalo A. Ruz and Matías Irarrázaval

Abstract: Early detection of suicide risk is crucial for schools and represents a significant public health challenge. This work presents a data-centric approach integrating psychological scales, domain expertise, and machine learning. The method: (i) combines traditional assessments with AI, (ii) follows a multi-model, data-centric AI strategy, and (iii) employs recurrence and feature extraction for high-fidelity monitoring. By prioritizing high-quality data and adaptive retrieval, it enhances generalization across contexts. Preliminary results from 2023–2024 in Chile, based on ~1,000 students, show strong potential, especially in low- and middle-income countries.

#98: On Fairness Vulnerability: A Data Poisoning Attack Perspective

Authors: Eunice Chan and Hanghang Tong

Abstract: With the growing adoption of AI and machine learning systems in real-world applications, ensuring their fairness has become increasingly critical. The majority of the work in algorithmic fairness focus on assessing and improving the fairness of machine learning systems. There is relatively little research on fairness vulnerability, i.e., how a system's fairness can be intentionally compromised. In this work, we introduce a simple yet effective attack designed to compromise fairness with little impact on the overall accuracy. We propose a novel attack on fairness that can be applied to any data-driven classification system, achieving a stronger fairness-accuracy trade-off compared to existing methods.

#101: The Challenge of Privacy in Assessment Data

Authors: Nancy Le, Alison Cheng and Nuno Moniz

Problem Statement: The proposal addressed the pervasive issue of data privacy in educational assessments, particularly in the context of increasing reliance on AI technologies. The risks associated with processing and utilizing sensitive student data are significant, especially when considering the involvement of minors who often lack control over their personal information (Alim et al., 2018). The proposal argues that existing datasets in educational contexts not only pose risks of unique identification of individuals but also highlight systemic weaknesses in safeguarding student privacy amidst the expansion of data-driven educational practices (Reidenberg et al., 2018).

AI Innovation: In response to the identified privacy risks, the proposal employed advanced analytical methods to examine publicly available educational assessment datasets (Realinho et al.,2021). The innovative approach involved assessing the effectiveness of various privacy-preserving techniques such as anonymization, perturbation, and synthetic data generation. By thoroughly analyzing five different datasets, we discovered that conventional privacy measures were insufficient in mitigating the risks of single-out occurrences (Stahl & Karger, 2016). This finding revealed an urgent need for more robust and sophisticated privacy preservation strategies that ensure the integrity of student data while enabling effective AI-driven educational assessment technologies.

Translational Evidence: The proposal discusses the broader implications of the findings for policy and practice. The results emphasize the necessity for educational institutions and policymakers to re-evaluate current data protection methods and legislation (Stahl and Karger, 2016). There is significant potential for deploying improved privacy-preserving practices in educational assessment data management, leading to enhanced trust and security for students and their families. By developing and implementing comprehensive policies that prioritize student privacy, educational institutions can facilitate more secure and responsible utilization of data-driven AI solutions, ultimately fostering a safer learning environment and promoting equity in educational opportunities.

RISE Dimensions: This innovation aligns with several RISE dimensions, notably Responsibility, Safety, and Ethics: Responsibility: Educational institutions hold a fundamental obligation to protect the sensitive data of students, particularly minors. The proposal advocates for the adoption of responsible data management practices that prioritize student privacy. Safety: With heightened awareness of the risks associated with data breaches, the implementation of advanced privacy measures enhances the overall safety of student information within educational environments. The findings call for a proactive approach to data collection and usage, ensuring that student data is adequately secured against potential harm. Ethics: Ethical considerations are paramount in data-driven contexts, particularly when dealing with vulnerable populations like students. The chapter highlights the need for ethical frameworks to guide the use of AI and data in education, ensuring that technology is leveraged for positive outcomes while minimizing risks of exploitation or harm. In conclusion, this proposal underscores the significance of addressing data privacy challenges in educational assessments through innovative AI methodologies. By advocating for responsible practices and ethical considerations, the authors contribute to the discourse on safe and equitable educational environments, reinforcing the imperative for protective measures that respect student agency and privacy.

References:

- Alim, F., Cardozo, N., Gebhart, G., Gullo, K., & Kalia, A. (2018). Spying on Students: School-Issued Devices and Student Privacy. Electronic Frontier Foundation Link.
- Reidenberg, J. R., & Schaub, F. (2018). Achieving Big Data Privacy in Education." Theory and Research in Education, 16(3), 263–279. Link.
- Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students' Dropout and Academic Success [Data set]. UCI Machine Learning Repository. Link.
- Stahl, W. M., & Karger, J. (2016). Student Data Privacy, Digital Learning, and Special Education: Challenges at the Intersection of Policy and Practice. Journal of Special Education Leadership, 29(2), 79–88. Link.

#102: Utilizing AI to Enhance Law Enforcement Response and Public Safety in Underserved Communities

Authors: Abby Sabella and Annabel Brown

Abstract: Underserved communities across the United States, particularly in rural, underfunded, and racially diverse areas, face significant challenges in providing law enforcement due to officer shortages, slow response times, and retention issues. These challenges have created a system of inadequate law enforcement that has made inefficient public safety the new normal, where agencies are consistently overburdened and lack the staffing and support necessary to keep up with administrative tasks. This inefficiency leads to a lack of time for community interaction and officer training, which in turn facilitates police misconduct and public distrust of the police. This project will address how artificial intelligence can be utilized to fill these gaps, creating more effective and faster law enforcement responses that streamline public safety.

Throughout this process, we will examine how innovative tools such as predictive policing algorithms, AI gunshot detection systems, automated report generation, and body camera review technologies could significantly affect policing agencies. These tools have the potential to ease the burden of many administrative tasks that police officers are tasked with, allowing the allocation of police officers to more public-facing tasks, increasing community interaction and rebuilding trust and relationships with the public. While these emerging technologies show great potential, these tools are imperfect and raise ethical concerns surrounding transparency and civil liberties. This research will explore the crucial role of public policy and governing necessary to efficiently implement AI in law enforcement while protecting citizens' right to privacy. Transparency and civilian oversight are essential to ensuring the ethical use of these technologies, and we plan to include research on individual cities, such as Seattle, WA, that have already begun to use civilian oversight as a mechanism for ensuring effective and safe law enforcement technology usage.

Furthermore, these tools are not an all-encompassing fix to the problems faced by underserved law enforcement agencies, as every community faces unique circumstances and systemic barriers to the implementation of these technologies. The implications for increasing efficiency and public safety demonstrate the need to continue developing these tools, making them more affordable and accessible, and encouraging their utilization nationwide as we learn to address and circumvent the problems associated with them.

#119: Misguided by Machines and Minds: How AI Exploits Our Need to Belong and Fuels Misinformation

Authors: Matthew Facciani

Topic for panel or presentation focused on my forthcoming book, Misguided: A discussion into how AI amplifies misinformation not just through flawed outputs, but by exploiting fundamental human psychology—our need for identity, belonging, and certainty. This session connects insights from Misguided to the ethical, educational, and institutional challenges of living with AI in a fragmented information ecosystem.

Abstract for individual paper: As AI becomes embedded in everyday platforms—from social media to search engines to chatbots—it's reshaping how people find, trust, and share information. But our biggest vulnerability isn't the technology—it's us. Drawing from my book Misguided, this talk explores how AI systems are accelerating misinformation by aligning with deep psychological and social forces, especially our identity-driven biases. I'll show how our "Paleolithic emotions" and need for belonging can be hijacked by AI-generated content, and why the real challenge is not just technological error, but human susceptibility. We'll examine the double-edged role of AI: as a tool that can scale disinformation, but also as a potential ally in building resilience—through prebunking, digital literacy education, and better transparency. This session will argue that solutions must be as interdisciplinary as the problem itself: not just smarter tech, but also stronger institutions, better education, and a deeper public understanding of our own minds. If we hope to navigate the future of misinformation, AI literacy must evolve hand-in-hand with media literacy, civic trust, and psychological insight.

#131: Implementing Artificial Intelligence in the Hospitality Sector to Combat Trafficking of Children

Authors: Monalisa Ms and Saranya Chandrasekaran

Abstract: This paper explores the role Artificial Intelligence (AI) as applied in the hospitality industry to combat child sex trafficking (CST). The paper looks at the current adoption of AI in the hotel industry, its specific usage to address trafficking and the possibility of curating AI tools to further enhance the safety of children. Human trafficking is "one of the most profitable and violent forms of international crime, generating an estimated $150 billion worldwide per year". In 2022, the International Labor Organisation (ILO) estimated that in 2022, approximately 2 million children were in commercial sexual exploitation. The hospitality industry has inadvertently contributed to the problem of CST.

Hotels are consistently used by child sex traffickers as the site of crime. Hotels are used to provide temporary shelter to victims as well as a location for abusing children. Hotels are also pit stops for sale and exchange of children or removing them from the care of their guardians. The hospitality industry has experienced legal ramifications for its role.

Statutory obligations as well as lawsuits have followed the industry in the US. It has grappled with dozens of lawsuits filed by victims in various courts across the United States. The claims allege a range of actions and omissions of the industry from failing to stop the abuse to actively profiting from it. This is a critical financial risk for an economy where travel bookings contributed $1.7 trillion in 2017 alone. Further risks arise as lawmakers contemplate more stringent statutory requirements for hotels in particular.

For instance, 12 states require hotels to train their employees on identifying signs of trafficking in their premises. The hospitality industry has attempted to address the problem with the traditional route of training its staff and updating its policies yet it has not made a significant dent in the problem. Now new age hotel managers are turning to technology- and Artificial Intelligence- to address the problem as well as limit their legal liabilities. Some ways in which the industry has already adopted technology led solutions include using maps that show smuggling and trafficking routes and pinpoint the properties in their portfolio close to them, in order to prevent human trafficking. Further, an increasing number of managers use criminal heat maps displaying properties' various levels of risk per category. Hotels also use AI for smoother day to day functioning. This paper argues that AI deployment for routine functioning can be curated for safeguarding children. This is in keeping with the statutory requirement of training staff and in the spirit of taking impactful action. The paper proposes specific examples of AI usage including:

1. AI led predictive maintenance- The classic example is that of an unoccupied hotel room logging low temperature which draws the attention of the maintenance team to fix a faulty radiator. This allows proactive scheduling, uninterrupted revenue from the room and reduction of breakdown time and costs. This usage can be curated to flag rooms that don't allow cleaning for days- a classic sign of a hotel room being used for child sex trafficking. Not allowing staff access to a room is also a red flag that hospitality staff is trained to identify.
2. AI led internet scanning- It is common practice for hotels to closely follow their media coverage. Massive media teams and other resources are dedicated to constantly monitor the brand presence and profile. Simultaneously, traffickers typically use social media and specific websites to share details of their services, availability of children, and location- specifying hotels. Currently hospitality are being encouraged to use AI led tracking of the Internet for any advertisements from sexual services sites that mention their hotels.
3. AI curated customer experience- Hotels currently rely on AI to crunch guest data, preferences and habits to personalize their hospitality experience. Further, analysis of historical data can forecast customer demand as well as ensure customer retention. Curation of such AI usage can also safeguard children: for instance, staff are trained to identify problematic behavior of guests and red flags associated with traffickers. Thus, an AI enabled analysis of guest behavior against identified red flags can equip hotels with critical information, safeguard vulnerable children, cooperate effectively with law enforcement as well as a way to avoid legal liability for failure to act. The paper argues these and other examples of AI led interventions can ensure that traffickers' ecosystem of crime can be interrupted to safeguard vulnerable children. It can also effectively lead the model of AI intervention in spaces and industries that have become complicit in CST.

Track: Education and Workforce of the Future

#7: Advancing AI Capabilities and Evolving Labor Outcomes

Authors: Yong Lee and Jacob Dominski

Abstract: The rapid advancement of Generative AI in recent years has intensified concerns about job displacement. Initially, these concerns focused on predictive tasks where Machine Learning excelled. However, more recently, attention has shifted toward occupations that involve editing, writing, coding, and content creation. Generative AI's progress is outpacing how organizations and workers are adapting, leaving many struggling to figure out and worried about how best to incorporate these technologies into their operations, workforce management, and organizational structures. While the potential for AI to displace workers, reshape tasks, and reorganize firms has generated much hype and concern, we still know relatively little about its current impact, and uncertainty looms about its future effects. Generative AI is a very new technology, and it is difficult to quantify how organizations and individuals are actually using it. Data on AI adoption is limited, and existing efforts to collect data lag behind the fast pace of AI development. Moreover, by the time such data is published, it often reflects outdated AI model capabilities.

To design better policies, training programs, and frameworks for AI adoption, we need near real-time insights into AI's effects on the labor market. This project aims to address this knowledge gap by examining the exposure of various occupations to state-of-the-art Generative AI models and exploring AI's near real-time impact on employment levels, unemployment rates, work hours, and labor demand. To assess the exposure of different occupations to state-of-the-art Generative AI models, I use prompt engineering to query various large language models (e.g., ChatGPT, Llama, Claude, etc.). I query the percentage of tasks within each tasks that can be effectively handled by these AI systems. Task data is sourced from the Bureau of Labor Statistics' O*Net dataset, which provides detailed information about U.S. occupations, the specific tasks associated with each role, and the relevance of those tasks. Using this task-level analysis, I construct an "Occupational Gen AI Exposure Score" to quantify the exposure of each occupation to AI. Additionally, I utilize the Current Population Survey (CPS), a monthly survey of approximately 60,000 U.S. households conducted by the Bureau of Labor Statistics, to examine trends in employment, unemployment, and other labor market indicators. By conducting quasi-experimental analyses, I compare high-exposure and low-exposure occupations before and after major AI model releases to estimate AI's impact on labor outcomes.

#18: Increasing AI Literacy and Appropriate Use in Nursing Education

Authors: Nicole Mentag and Jandra Antisdel

Abstract: Increasing AI Literacy and Appropriate Use in Nursing Education Dr. Nicole Mentag, PhD, RN and Dr. J'Andra Antisdel, PhD, RN Background and Problem Statement Nursing students with low AI literacy are at risk of experiencing consequences associated with misinformation, relying on AI-generated content without recognizing its inherent limitations [1,2]. Further, students may not learn how to engage with AI effectively to enhance their learning and clinical skills, risking further ramifications of improper or unethical use. Nursing educators are responsible for ensuring that students are informed about AI and must include AI in their pedagogical approach to increase AI Literacy in nursing students [3].

AI Innovation: Two applications of AI were used in nursing education: (1) an AI-generated writing activity designed to expose students to inaccuracies and errors in AI-generated content; and (2) an AI-generated voice simulation to give students experience in initiating sensitive patient assessments. In-Class AI Writing Activity In an undergraduate nursing research course, an in-class activity was designed to apply their knowledge of APA style with an AI-generated scholarly paragraph and reference list. During the activity, students initially focused on identifying APA errors. However, as they explored further, they found inconsistencies in the AI-generated content, such as hallucinated references and broken hyperlinks. The discovery of these issues led to a robust class discussion, addressing the implications of using AI and the risks of relying on AI without critical analysis. Students were shocked to learn that the paragraph and reference list were generated by GPT-4 [4], raising concerns about the reliability and potential consequences of using AI tools in their professional writing and nursing practice. This in-class activity offered students a chance to practice APA formatting. It

prompted a larger conversation about the potential consequences of relying on AI tools, from the dangers of misinformation to the reinforcement of biases. This conversation evolved into strategies to incorporate AI into professional writing while maintaining the integrity of nursing research and practice. AI Voice Simulation for Sensitive Patient Assessments During the orientation of their mental health clinical rotation, Nursing students participated in a discussion about their concerns regarding psychiatric nursing and inpatient mental health settings. The most common anxieties expressed were talking to patients and asking sensitive questions, particularly about suicidal ideation and depressive symptoms. Following this discussion, students were introduced to an AI-Voice simulation using GPT-4 [4]. GPT-4 was prompted to act and respond as a standardized patient, a male patient experiencing marital and financial issues and who had been struggling with depressive symptoms. Each student took turns engaging with GPT-4 by practicing open-ended questions and therapeutic communication. By interacting with an AI-powered simulation, students could practice without judgment, receive instant feedback, and build confidence. After the exercise, students reported feeling more prepared and less anxious about asking sensitive questions. This activity reinforced therapeutic communication skills before students encountered real patients.

Translational Evidence & RISE Dimensions: These innovative learning activities are congruent with the guiding principles of RISE, addressing the dimensions of responsibility, inclusion, safety, and ethics. They provide students with opportunities to engage with AI, which recognizes and addresses the responsibility of nurse educators to prepare nursing students to use technology in their future roles as healthcare leaders. The use of AI for education can be an equalizer for low-resource institutions, allowing for inclusion and increasing access. As an example, simulation training in suicide has been used to increase skills, knowledge, and behaviors in healthcare professionals [5]. However, traditional methods such as the use of standardized patients, virtual reality, and high-fidelity mannequins can be limited by cost, availability, and access, especially in more underserved areas [6]. AI is currently used in healthcare to support diagnostics and clinical decision-making3 and unfortunately, there are consequences when biases are not recognized. For example, an algorithm that used race, as a factor in patient care, led to unsafe practices and poor and inequitable outcomes in people of color [7]. Given the consequences of unreliable and biased AI-generated content, these learning activities can provide nursing students with the basic understanding of AI. Students need to build the skills to recognize ethical concerns such as potential inaccuracies and biases when engaging with AI in the healthcare setting. By embedding these AI innovations in nursing education, students will be empowered to use AI effectively while critically evaluating its limitations, fostering responsible, inclusive, safe, and ethically informed patient care practices.

References:

1. De Gagne JC, Hwang H, Jung D. Cyberethics in nursing education: Ethical implications of artificial intelligence. Nursing Ethics. 2024;31(6):1021-1030. doi:10.1177/09697330231201901
2. Khan B, Fatima H, Qureshi A, et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomedical Materials & Devices. 2023;1(2):731-738. doi:10.1007/s44174-023-00063-2
3. American Nurses Association, ed. Code of Ethics for Nurses. 2025th ed. American Nurses Association; 2025.
4. OpenAI. ChatGPT [Large language model]. Published online 2023. https://chat.openai.com/chat
5. Richard O, Jollant F, Billon G, Attoe C, Vodovar D, Piot MA. Simulation training in suicide risk assessment and intervention: a systematic review and meta-analysis. Med Educ Online. 28(1):2199469. doi:10.1080/10872981.2023.2199469
6. American Nurses Association Center for Ethics and Human Rights. The Ethical Use of Artificial Intelligence in Nursing Practice.; 2022. Accessed March 6, 2025. https://www.nursingworld.org/globalassets/practiceandpolicy/nursing-excellence/ana-position-statements/the-ethical-use-of-artificial-intelligence-in-nursing-practice_bod-approved-12_20_22.pdf
7. Gutiérrez OM, Sang Y, Grams ME, et al. Association of estimated GFR calculated using race-free equations with kidney failure and mortality by black vs non-black race. JAMA. 2022;327(23):2306-2316. doi:10.1001/jama.2022.8801

#23: Comparing Statistical and Machine Learning Methods in Missing Data Imputation in Graded Response Model

Authors: Yilin Li, Ke-Hai Yuan and Ying Alison Cheng

Abstract: Missing data is a pervasive issue in research across various fields, posing significant challenges to the validity and reliability of conclusions drawn from incomplete datasets. This problem is particularly pronounced in fields like psychology, education, and health sciences, where survey and assessment data often include missing item responses. Ignoring or inadequately addressing missing data can lead to biased parameter estimates, reduced statistical power, and compromised generalizability of findings. Addressing missing data issues effectively is therefore a crucial aspect of ensuring the integrity of research findings. Researchers have developed numerous methods for handling missing data, each with distinct theoretical underpinnings, advantages, and limitations. Ad-hoc methods include mean substitution, and deletion methods such as listwise deletion and pairwise deletion, which are straightforward but can lead to substantial loss of information and biased estimates. Principled approaches, such as multiple imputation (MI) and maximum likelihood estimation (MLE), have been widely embraced for their flexibility under the assumption of data missing at random. While MI involves generating multiple plausible datasets and combining results to account for uncertainty, MLE focuses on deriving parameter estimates by maximizing the likelihood function given the observed data. Despite their advantages, these methods can be computationally intensive and rely on correct specification of the underlying data model. Recent advancements in computational methodologies have enabled the adoption of machine learning techniques for missing data imputation. Decision trees and random forests, for instance, leverage the predictive power of ensemble learning to estimate missing values by modeling complex relationships within the data.

Other studies have explored neural-network-based imputation techniques for their adaptability in capturing complex data patterns, particularly in non-linear and high-dimensional contexts. Graded response models (GRM) are used to measure a latent trait with polytomous response categories. Emerging evidence suggests that the choice of an imputation technique may also depend on specific aspects of GRMs, such as the ability to maintain the accuracy of item information functions and satisfy model assumptions, including unidimensionality and local independence. These considerations highlight the need for targeted evaluations of imputation approaches under psychometric frameworks. Despite the variety of imputation methods available, the choice of an appropriate technique often depends on the nature of the data, the extent of missingness, and the specific goals of the analysis.

The primary aim of this study is to compare the accuracy of different imputation methods, including ad-hoc approaches, MI, the EM algorithm, and machine learning-based techniques such as decision trees and random forests, in the context of GRMs. A Monte Carlo simulation will be conducted to evaluate the performances of several commonly used missing data imputation methods with GRM, in terms of parameter recovery, model fit, and robustness. In addition, an empirical data set will be employed for illustration purposes. By employing an empirical dataset, we aim to evaluate the performance of these methods in recovering missing responses and maintaining the integrity of psychometric properties. The simulation aims to answer this question: What are the effects of missing data mechanisms, the proportions of missing data, the number of items, and the number of response categories on the performance of missing data imputation methods in GRM? We will consider the following design factors: (a) missing data mechanisms, including missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR); (b) the proportions of missing data of 5%, 10%, 20%, 30%, or 40%; (c) the number of items of 10, 20, 30, or 50; (d) the number of response categories of 3, 5, or 7. Complete data sets are created in GRM using R. Missing data will be simulated according to the Bernoulli distribution where the parameter is modeled by a logistic regression. The translational impact of this research is relevant to social science research, where self-report questionnaires and psychometric assessments are frequently used to measure latent constructs such as anxiety, depression, and personality traits. Missing responses in these assessments, often due to participant fatigue, response omissions, or sensitivity to certain items, can significantly impact the validity of psychological studies.

For example, in longitudinal studies of mental health, missing data on key symptom measures may distort trajectories of psychological well-being over time, leading to incorrect conclusions about treatment efficacy or the progression of disorders. Similarly, in large-scale personality assessments, improper handling of missing responses can alter factor structures and misrepresent individual differences, ultimately affecting the validity of psychological theories and

practical applications in clinical and organizational settings. By integrating AI-driven imputation techniques, this study provides a robust solution for recovering missing data while preserving the integrity of psychometric properties, ensuring that psychological research produces more reliable and generalizable findings. This study aligns with the responsibility dimension of RISE by emphasizing the development and evaluation of imputation methods that uphold the integrity and reliability of research findings. Responsible data handling is particularly crucial in high-stakes assessments, such as educational testing and clinical research, where missing responses can influence policy decisions and individual outcomes. For example, in large-scale standardized testing, an inadequate imputation method may introduce biases that unfairly advantage or disadvantage certain groups of students, leading to inaccurate inferences about educational performance gaps. Similarly, in health research, improper imputation of missing clinical data may distort patient risk assessments, potentially leading to inappropriate treatment recommendations. By systematically evaluating AI-based imputation techniques within a psychometric framework, this study ensures that missing data is addressed with transparency, fairness, and methodological rigor, fostering responsible research practices that enhance the credibility of empirical findings.

#44: AI-Enhanced Reflective Practice: Increasing Access to Faculty Development

Authors: Eric de Araujo

Abstract: Problem Statement While doctoral programs excel at producing disciplinary experts, they are not, in general, proficient in developing instructors of that discipline. As a result, faculty are often insufficiently skilled in helping students traverse the path from novice to expert. Institutional faculty development efforts serve to remediate this defect of doctoral training. Faculty at well-resourced institutions can benefit from working with experts to improve their teaching skills through frameworks like Decoding the Disciplines, which make explicit the disciplinary assumptions that help students overcome bottlenecks on their journey to mastery. However, many faculty, particularly contingent faculty at under-resourced institutions, lack access to such development opportunities, creating a significant equity gap in higher education.

AI Innovation: I present two AI-powered tools that, in the absence of dedicated faculty support, guide faculty through parts of the Decoding the Disciplines framework. The first tool engages faculty in a structured dialogue to identify what the framework calls a "bottleneck"—a specific point where students struggle to grasp disciplinary ways of thinking. The second tool generates a visual concept map of this bottleneck, helping faculty appreciate the latent knowledge and skills their novice students need to master this threshold concept. Together, these tools promote reflective teaching practices without requiring the presence of a faculty development expert.

Translational Evidence: An early presentation of these tools at a regional teaching conference received positive feedback, demonstrating their potential to help educators independently identify and address discipline-specific bottlenecks. Unlike generic AI assistants, these purpose-built tools embed knowledge of a well-researched educational framework through targeted prompting strategies. Moreover, this approach to tool development is not specific to this particular educational framework, but represents a template for how AI can provide pedagogical support to faculty across various contexts. This proof of concept demonstrates the scalable potential for addressing faculty development needs where institutional resources are lacking.

RISE Dimensions Inclusion: By reducing barriers to implementing sophisticated educational frameworks, these tools democratize access to high-quality faculty development resources. This particularly benefits educators at resource-constrained institutions and in under-served regions where faculty development opportunities are limited, addressing a significant equity gap in higher education. Responsibility: The tools are designed to augment rather than replace human reflection. They preserve educator agency by acting as interlocutors rather than simply providing recommendations. This approach embodies responsible AI by enhancing professional practice without automating it, as the tools prompt educators to think more deeply about their discipline-specific knowledge rather than generating educational content themselves. This work demonstrates how intentionally designed AI applications can address educational inequities by making sophisticated pedagogical frameworks accessible to all educators, while maintaining the critical human element of reflective practice in teaching.

#64: AI-powered, Targeted Instructional Support for Early Childhood Teachers

Authors: Jill Pentimonti, Toby Li, Michelle Luna, Sumin Hong, Rick Johnson and Tricia Zucker

Abstract: AI-powered, Targeted Instructional Support for Early Childhood Teachers RISE AI Conference Proposal Problem Statement: This exploratory study aims to establish proof of concept for an AI-powered instructional support system for early childhood teachers. The long-term focus of the support system will be on teacher behaviors that support children's development of language, literacy, and STEM skills. Supporting the development of these skills early is essential for children's future academic success, particularly for those children from vulnerable populations who are at-risk for later learning difficulties (Morgan et al., 2016). This untested and potentially transformative instructional support system will build on our work developing the Systematic Assessment of Book Reading (SABR; Pentimonti et al., 2021), a tool that examines qualities of teacher talk through observation of classroom shared book reading and is predictive of children's achievement in critical early learning skills. The SABR tool requires low-inference, keyword coding by trained members of a research team. The current, necessary human coding time for SABR is significantly less than other high-inference, observational measures (e.g., CLASS is 40 hours training/certification plus 2-4 hours per observation) but is still substantial (i.e., 8 hours of training/certification plus 0.5 hour video coding per observation), which presents practical challenges to providing timely feedback to teachers at scale. Yet SABR's low-inference, keyword coding format makes it an ideal process to convert to coding via a Natural Language Processing (NLP) pipeline and has the potential to promote more brisk and lower-cost feedback to teachers on instructional targets known to improve children's learning.

Our investigative is first developing machine learning (ML) models necessary to apply SABR codes via NLP, as opposed to human coding. Next, we are engaging a focus group of teachers in the user-centered design process (Vredenburg et al., 2002) in order to establish a prototype of an innovative AI-powered app that would share feedback with teachers based on SABR scores in a timely manner. This process also involves exploring ways to provide teachers with access via the app to professional development (PD) materials aligned with their SABR-identified instructional learning needs. Importantly, this work is designed to support teachers in under-resourced classrooms who typically have limited access to costly PD. The technology we are developing will ultimately provide a cost-effective method to receive evidence-based PD to budget-limited early childhood programs. Additionally, the use of emerging natural language processing and large language model (LLM) technology (Tan & Jiang, 2023) brings the benefits of these new artificial intelligence (AI)-powered technologies within reach of education professionals.

AI Innovation: This project is exploratory and transformative, as it represents a radically different approach to the provision of timely feedback to teachers on instructional practices, as well as the first attempt to use emerging AI technologies as a means to provide content-focused PD in an efficient, scalable, low-cost manner in early childhood classrooms. This project is also makng significant technological contributions at the intersection of Human-Computer Interaction (HCI) and NLP. Specifically, it is creating (1) a new NLP pipeline that leverages the strengths of state-of-the-art LLMs and the rigid structures of the SABR framework to generate timely, specific, and actionable feedback to teachers during book reading in early childhood classrooms; and (2) exploring novel interfaces that effectively deliver the AI-generated feedback to teachers in formats that are appropriate for the context. Furthermore, this project is contributing to the knowledge base concerning approaches to instruction that utilize innovative technology solutions to deliver personalized, timely pedagogical support, particularly for under-resourced classrooms. This work constitutes the next vertical area of research in this space, because prior work has established the instructional behaviors measured by the SABR as those with the most impact on foundational learning skills. However, investigations of innovative methods for providing feedback and support to teachers in utilizing these instructional behaviors is lacking.

Translational Evidence: This project will have critical societal benefits as findings inform recommendations for practitioners and policymakers about the value of investing in technological solutions to provide pedagogical support to teachers, particularly those working in under-resourced schools. Further, this work will illuminate how technological support might address the lack of PD resources for teachers in low-resourced schools and achievement gaps for young children from low-income families.

RISE Dimensions: This project aligns with the RISE dimensions of responsibility, inclusion, and safety. First, we ensure that our engagement in the design process for the AI-app is focused on responsibility – specifically, as we develop the app we take into account any consequences of the AI systems for both teachers and children in their classrooms. Second, when engaging with teachers in the development process we are sure to incorporate inclusive practices and draw on perspectives from a diverse group of teachers to help us identify potential ethical concerns and collective efforts to address those concerns. Finally, the project works to ensure safety for our participants through activities such as keeping the identity of children protected and keeping original recordings private.

#71: Teacher Simulator: Synthetic Dialogue Generation for AI-Powered Instructor Development

Authors: Isabel Molnar, Si Chen, Khiem Le, Ting Hua, Ron Metoyer and Nitesh Chawla

Problem Statement: Large Language Models (LLMs) hold promise for supporting educator development, yet they struggle with multi-turn scaffolding—the ability to iteratively refine instructional guidance over multiple conversational turns. Current AI-powered tutors primarily focus on student learning and content delivery, often failing to provide structured, adaptive coaching for teachers themselves. Effective teacher development requires high-quality, structured dialogues that capture novice-instructor challenges, expert feedback strategies, and pedagogical reasoning over extended interactions. However, collecting real-world conversational datasets is challenging due to data scarcity, privacy constraints, and pedagogical complexity. Without high-quality training data, AI-generated teaching support remains rigid, unstructured, and detached from real-world classroom decision-making.

AI Innovation: To address these challenges, we propose a Teacher Simulator, an LLM-powered framework that synthesizes structured pedagogical dialogues by interacting with human experts. Unlike traditional AI tutors that emphasize student problem-solving, our simulator is designed to model teacher learning, instructional decision-making, and iterative pedagogical refinement. Our contributions include: Simulating Novice Teacher Mistakes: Using LLM-driven behavioral modeling, the system generates common novice teacher errors (e.g., ineffective questioning, poor scaffolding, misinterpreting student responses), allowing pedagogy experts to diagnose and provide structured feedback. Co-Designed Instructional Frameworks: AI-generated conversations follow predefined, research-backed scaffolding structures, ensuring consistency and effectiveness rather than producing ad hoc, unstructured responses.

These frameworks incorporate Socratic questioning models, instructional adaptation strategies, and error diagnosis techniques aligned with evidence-based pedagogy. Multi-Turn Adaptive Scaffolding: Unlike traditional AI one-shot recommendations, our system sustains complex multi-turn interactions, dynamically adjusting its suggestions based on instructor responses. The AI refines teaching strategies over time using reinforcement learning, real-time feedback loops, and progressive difficulty scaling. Human-in-the-Loop Validation: To ensure pedagogical accuracy, bias mitigation, and inclusivity, our approach integrates expert educator validation, bias audits to detect overrepresentation of certain teaching styles, and accessibility reviews to support diverse learner needs. By combining synthetic data generation with expert-informed instructional strategies, our approach creates a scalable AI-driven instructor development framework that goes beyond content delivery to deep pedagogical support.

Translational Evidence & Socio-Economic Impact: Our Teacher Simulator has the potential to: Improve novice teacher training, particularly in under-resourced institutions where expert mentorship is limited. Facilitate interdisciplinary instructional support, ensuring AI-driven teaching guidance is adaptable across STEM, humanities, and special education. Enhance responsible AI adoption in education, ensuring AI amplifies, rather than replaces, evidence-based teaching strategies. Scale high-quality pedagogical mentorship, enabling AI-powered tools to address educational disparities and support faculty in diverse institutional settings.

#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.

#95: AI and the Future of Executive Leadership: Evidence-Based Insights on Decision-Making, Skills Development and Workforce Engagement from Executive MBA Students in Pan African Business School

Authors: Joseph Onyango

Abstract: AI and the Future of Executive Leadership: Insights on Decision-Making Efficiency, Skills Development and Worker Engagement Among Executive MBA Students from the Pan African Business School Background Artificial Intelligence (AI) integration is revolutionizing executive leadership and decision-making across several industries. Organisations are grappling with the challenge of upskilling their leaders with the skills to succeed in AI-enhanced workplaces as AI technologies grow. With a focus on the socio-economic ramifications of well-honed AI implementation, particularly to emerging markets, this research focus on the principal competencies required for competent executive leadership within the AI domain. To thoroughly investigate these issues, the study adopts a mixed-methods methodology, combining quantitative and qualitative methods.

This study aims to provide: (1) insight into the implications of developments in AI for executive roles and decision making in organisations in multiple industries; (2) identification of critical new skillsets that executives and professionals will need for success in an AI-augmented workplace; (3) description of how organizations can proactively and strategically invest in upskilling and reskilling initiatives to ease AI adoption; and (4) exploration of frameworks that executives may use to balance efficiency gained through AI with the role of human labor in organizational success. This approach is derived from Signaling Theory, as the process of Executive Communication strongly drives both executive behaviors and organizational strategy. This presents the grand challenge of ensuring artificial intelligence is at the core of leadership and decision-making in the fourth industrial revolution organisations of the present and future. The emergence of AI technologies necessitates a thorough comprehension of the effects of AI advancement on executive positions, the skills needed to succeed, and the steps that should be taken to execute these AI-adopting strategies. Because of the increasing significance of AI, many leaders are apprehensive about prioritising technological efficiency over stimulating human collaborators and providing growth opportunities. This disconnect is risky for organizational performance, employee morale, and job security. However, we must understand the interdependence of AI technologies and the role of executive leadership in managing the complexity of today's contracts (Diao, 2024).

Methods: The study utilizes a mixed methods-based research design to examine the extent and implications of integration of AI into leadership. The methodological approach is centred on Signaling Theory. This theoretical proposition states that information diffusion within the structure of organizational behaviour and functioning significantly influences collective behaviours and decision-making (Kshetri, 2021). This study is based on two concrete parts: a quantitative survey and qualitative interviews. The quantitative aspect of the study comprises a structured Likert-scale survey administered to 56 Executive MBA students at Strathmore University, who represent a cross-section of future business leaders currently uniquely positioned to shape the way AI is adopted in their organisations. The survey seeks to gauge their understanding of AI, how embedded AI tools are in their organizations, and the perceived effects of AI on decision-making and leadership effectiveness. The qualitative portion consists of semi-structured interviews to collect data on in-depth perspectives on leaders, experiences, and recommendations around embedding AI into leadership. Thematic analysis identifies recurring themes in the qualitative data and can provide rich contextual insights that complement the quantitative findings (Yap, 2024).

Findings: This study produced a number of interesting results. We found a moderate positive correlation between executives, knowledge of AI systems, and AI integration into their organizations (r = 0.487, p = 0.006). It indicates that executives using AI tools tend to be in organizations that can meaningfully integrate it (Liu, 2024). The participants in the study identified other essential emerging work skills for success with AI addition: data literacy, adaptability, emotional intelligence and strategic decision-making. This blog post [is linked to the current prompt] highlights that treating their workforce this way is the secret to overcoming the complexities of AI integration and leading well in this new context (Farayola et al 2023). Organisations have acknowledged the importance of upskilling initiatives and have done them. Still, the gap exists that there is no correlation (p = 0.139) in improvement in executive decision-making skills in organisations with the initiatives taken. This reinforces the importance of organizations to invest in training strategies to fit leaders with positions for effective performance (Jobin et al., 2019). Open-ended responses provided qualitative insights that confirmed the need for a balance between AI efficiency and human labour. Leaders stressed the importance of facilitating a collaborative culture in which AI is an ally to human labor rather than a substitute, which is critical for sustainable growth (Xu & Zhang, 2023).

Implications: Translation Evidence Hence, the results of this research underscore necessary translational proof of the use of AI solutions and their likely socio-economic consequences. Organizations that adopt successful innovative organisations tend to use leaders who understand AI technologies. In addition, a positive correlation between AI assimilation and executives' better decision-making indicates that organizations, encapsulated by AI frameworks, will be able to not only expand performance indicators but also boost revenues with employee involvement (Khatib et al., 2024). Yet the study also highlights an important gap: the absence of a statistically significant correlation (p = 0.139) between investments in upskilling initiatives and improved leadership decision-making. This underscores the importance of organisations strategically investing in training initiatives designed around AI skills to fill identified knowledge gaps and ensure an agile working cohort. Developing emerging leadership skills is crucial because the capacity to respond to the complexities brought about by AI directly affects the success and resilience of organizations in an increasingly digitalized world (Mohan, 2024).

RISE Dimensions: The development and application of AI products and solutions in the context of this research aligns greatly with the RISE readability dimensions–Responsibility, Inclusion, Safety, Ethics. The study calls attention to the need for responsible use of AI in leadership positions, highlighting the ethical considerations involved in adopting AI. It encourages organizations to adopt approaches emphasising transparency and fairness to address biases potentially embedded in AI technologies, promoting ethical adherence and congruence with organizational principles. Such ethical oversight builds the trust that is needed between leadership and employees, which is critical in collaborative environments powered by AI (Cortellazzo et al., 2019). It then becomes a key finding, as the study finds that equitable access to AI training may be critical for success for all employees. This will go a long way in keeping the workforce engaged and minimizing feelings of alienation within the workforce, as long as the right segments of the workforce are trained and equipped with the necessary competencies. A balanced mindset—one where the workforce is supplemented, not substituted by AI—will continue to cultivate a workplace that values employee contributions while providing a feeling of safety to the staff, that is, fears of job loss due to replacement by AI could be addressed (Sipitanos et al., 2022). By analyzing the existing literature on AI influences on executive roles, the study offers practical insights that organizations can implement to address the challenges associated with AI deployment and also ensure that they maintain or achieve the socio-economic advantages from their technology infrastructure. policymakers and business leadership in developing frameworks that embody AI-expedited functionality while advocating for ethical leadership (Bekti et al., 2024).

Conclusion: This finding highlights the need to ensure that executive leaders are equipped with the skills to drive AI integration through. The results suggest that improving AI literacy in leaders would profoundly change the decision-making approaches and as a result elevate the overall performance of organizations. This mixed-methods study presents an integrated understanding of how organizations may configure their AI processes into leadership contexts, thus making informed decisions on AI-enabled transformation efforts. The findings serve as a call to action for organizations to align their human capital strategies with the rise of AI, helping to drive priority investments in training and development. Further studies still seem like ones to pursue, and it can expand on how AI could be composed with different kinds of leadership, whatever that may be over time.

Reference:

- Bekti, I., Sampurno, A., & Rahmadani, A. (2024). Industrial Revolution 4.0: a study of challenges for digital leadership in Indonesia. JEH.31 Volume 3, Issue 2: 35–41 Journal of Ecohumanism doi:10.62754/joe. v3i2. 5015 2019
- Cortellazzo, L., Cardoni, A. & Zampieri, G. The author has a stress on the role of leadership in a digitalized world: A review. Frontiers in Psychology, 10, 1593. doi:10.3389/fpsyg. 2019. 01593
- Diao, K. (2024). How Artificial Intelligence in the world of Project Management Forthcoming in Business Economics and Management doi:10.54097/23axpg43.
- FarayolaO, Alabi YA, Adeyemo A. (2023). AI technologies driven innovation business models: A review Computer Science & IT Research Journal, 2(1), 22-35. Doi:10.51594/csitrj. V2i1. 608
- Jobin, A., Ienca, M., & Andorno, R. (2019). International overview of AI ethics guidelines. Nature Machine Intelligence, 1 (9), 389–399. doi:10.1038/s42256-019-0088-2
- W.Khatib & R.Bader:(2024). Headline: AI TRiSM in Education: What it means for business executives — knowledge and decision making. International Journal of Theory of Organization and Practice (IJTOP), 3(1), 101-112. doi:10.54489/ijtop. v3i1. 290
- Kshetri, N. (2021). Developing use of Artificial Intelligence in human resource management in global south emerging economies: Some path-finding evidence. Management Research Review, 44(3), 399–409. doi:10.1108/mrr-03-2020-0168
- Liu, H. (2024). Artificial Intelligence and Economic Growth: Evidence from Industrial Structure Advances in economics management and political sciences, 1(1), 45–57. doi:10.54254/2754-1169/123/2024mur0143
- Mohan, R. (2024). Practicing AI within your organization: Building new capabilities for skills of success and humanities conference. Journal of Education and Science, 24(1), 135-150. doi:10.52783/jes. 2935
- Sipitanos, T., Abidin, Z., & Jamea, A. (2022). Action research and critical discourse analysis: A pathway to school principals' professional development Nontraditional students faced with obstacles in their quest for master's degrees in the university: the university and students in the digital age. doi:10.24940/theijhss/2022/v10/i6/hs2204-017.
- Xu, Y., & Zhang, L. (2023). Will scholar-type CEO sales scholars contribute more to the industrial AI transition of manufacturing companies? Industrial Management & Data Systems, 123(5), 887-903. doi:10.1108/imds-11-2022-0672
- Yap, S. S., & Singh, A. (2024). How FIKR personality assessment tool can help include emotional intelligence, achievement orientation, analytical thinking, and structured leadership in AI based business leadership. Bonview Journal of Contemporary Business Affairs, 42(6), 45-57. doi: 10.47852/bonviewjcbar42024142

#157: Bridging the LLM Accessibility Divide? Performance, Fairness, and Cost of Closed versus Open LLMs for Automated Essay Scoring

Authors: Kezia Oketch, John Lalor and Ahmed Abbasi

Problem Statement: Large language models (LLMs) have dramatically advanced natural language processing (NLP) capabilities across a wide range of applications, including education, healthcare, law, and government services. However, access to high-performing LLMs is increasingly mediated through commercial APIs, usage caps, and proprietary licensing. Closed models like GPT-4 and GPT-3.5 dominate the landscape but present significant accessibility challenges. These models are expensive to operate, require persistent internet access, and lack transparency in training data, architecture, and evaluation methodology. For researchers, developers, and public institutions in low-resource environments, these barriers are not trivial. They limit the ability to deploy LLMs at scale, conduct independent audits, or fine-tune models to local contexts. In contrast, the rise of open and open-source LLMs such as Llama 3, Qwen2.5, and OLMo, offers a potential pathway to mitigate these barriers by prioritizing affordability, modifiability, and equitable access. Despite this promise, the performance and fairness of open models compared to closed systems remains underexplored, especially in complex, high-stakes tasks like automated essay scoring (AES). AES is widely used in education for formative feedback, placement testing, and exam grading. It requires a deep understanding of argument structure, grammar, coherence, and relevance to prompts. Capabilities that are hard to achieve without high-quality LLMs. Moreover, biases in AES can lead to real-world consequences for students and applicants. Thus, determining whether open models can match the performance and fairness of closed models in this domain is critical to evaluating their real-world viability. This study addresses this challenge by presenting a comprehensive, multi-dimensional benchmark of nine leading LLMs in AES tasks, grounded in a framework of Responsible, Inclusive, Safe, and Ethical (RISE) AI development.

AI Innovation: We introduce a rigorous benchmarking framework that evaluates nine LLMs across multiple dimensions: performance, fairness, generation quality, and cost. These models span three categories: closed (GPT-3.5, GPT-4, GPT-4o), open (Qwen2.5-72B, Llama 2-70B, Llama 3-70B, Llama 3.1-405B, DeepSeek-R1), and open-source (OLMo 2-13B). The models were evaluated on two widely-used AES datasets: Student Assessment Prize (ASAP) and Cambridge Learner Corpus-First Certificate in English (FCE), containing over 15,000 human-scored essays across a range of prompts, including narrative, argumentative, and descriptive genres. ASAP provides holistic scores based on rubric-aligned dimensions, while FCE offers both overall and analytical scores from Cambridge English assessments. The AES task was evaluated under zero-shot and few-shot prompting conditions. In the few-shot setting, each model received three reference essays, scored low, medium, and high along with a new essay to be scored. Essays were scored by each model using their native API or instruction-following interface. Scores were normalized to a 0–1 scale to enable comparison across models and prompts. We ensured prompt consistency, input formatting alignment, and standardized output parsing to ensure comparability. We used five standard metrics for performance evaluation: Mean Squared Error (MSE), Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC), Spearman Rank Correlation (SRC), and Quadratic Weighted Kappa (QWK). GPT-4 achieved the best average performance across metrics, but the gap with top open models was narrow. For instance, Qwen2.5-72B achieved lower MAE (0.159) and MSE (0.045) than GPT-4 in the few-shot condition. Llama 3-70B outperformed GPT-4 on QWK (0.970 vs. 0.964), suggesting that these models are nearly indistinguishable from closed models in downstream performance under carefully tuned prompting conditions. To evaluate fairness, we implemented a split-plot ANOVA design and computed error disparities (Δ = human score – model score) across demographic subgroups in the FCE dataset. Independent variables included prompt type, prompt ID, model, age group, and race (Asian vs. non-Asian). Most models showed subgroup bias under 5%, with Qwen2.5 and Llama 3.1-405B exhibiting comparable fairness to GPT-4. We further examined interaction effects and found that prompt-specific bias often explained more variance than model-level differences, indicating that domain and genre influence bias emergence. We also analyzed the generation quality of each model. Using 68 standardized prompts from the two datasets, we had each model generate a corpus of essays. These generated essays were then scored by every other model and also compared against human-written counterparts. We visualized semantic similarity using t-SNE based on document transformer embeddings, revealing that machine generated text shows high semantic and structural similarity but is very distinct from human generated text. A three-way ANOVA confirmed significant differences in scores assigned to generated essays depending on both model origin and scoring model. Cost analysis was conducted at the token level. Using pricing data from OpenAI, DeepInfra, and Replicate, we computed average costs per essay prompt. Results show that open LLMs like Llama 3 are up to 37× cheaper.

Translational Evidence: The results demonstrate that open models are not only competitive but in many cases, viable alternatives to closed models in real-world settings. The implications span multiple domains: Education and Assessment: LLMs are increasingly used for grading assistance, formative feedback, and high-stakes exam scoring. Our results show that open models can offer performance comparable to closed models at drastically lower cost. This makes LLM-based AES financially feasible for public school systems, examination boards, and non-profit education providers. Moreover, models like Llama 3 and OLMo 2 can be run locally, offering additional advantages in data control and latency reduction. Government and Civil Services: Many governments process thousands of written responses annually in immigration, public feedback, and social service applications. Using auditable open models enables these agencies to automate processing with transparency. Llama 3's local deployment capabilities support offline operation, which is valuable in countries with unstable connectivity or strict data localization laws. Independent audits of AES systems using released prompts and scoring metadata can reduce the risk of algorithmic opacity in public decision-making.

Research and Low-Resource NLP: Low and Middle-Income Countries (LMICs) and institutions can adopt open and open-source models to conduct NLP research, evaluate regional language performance, and build tailored systems. For instance, OLMo's open weights and documentation allow training extensions in code-mixed or indigenous language contexts. Even if OLMo lags behind GPT-4 in performance, its modifiability makes it a useful foundation for experimental research and AI pedagogy. Teaching and Capacity Building: Open models facilitate NLP education, allowing instructors to demonstrate fairness audits, prompt engineering, and few-shot tuning with reproducible tools. Unlike black-box APIs, open models provide consistent behavior over time, which is essential for reproducible coursework. Educational institutions can use these models to teach students how to build, evaluate, and govern AI systems responsibly. Policy and Ecosystem Impact: Governments and international organizations designing national AI policies can use this evidence to justify procurement of open models. Cost, equity, and auditability are now measurable and can be considered alongside raw performance in AI adoption roadmaps. We recommend public investment in open-model infrastructure, including compute grants and regulatory sandboxes for open deployment trials.

RISE Dimensions: Our work operationalizes Responsible, Inclusive, Safe, and Ethical AI as follows: Responsibility: We release full model scores, prompts, generation data, and evaluation scripts to support reproducibility. Our error modeling framework captures where model bias occurs and how prompt structure, demographic attributes, and model architecture interact. Open-source evaluations allow institutions to replicate fairness tests or benchmark new models. Inclusion: We assess model fairness across demographic subgroups (e.g., race, age), prompt types (narrative, argumentative, commentary, etc), and response sources (human vs. LLM-generated). Open models facilitate inclusion by enabling researchers from underrepresented regions to study and improve model performance without license restrictions or API dependencies. We show that fairness and accuracy are not exclusive to closed models. Safety: AES systems directly impact educational outcomes. Bias in these systems can reinforce structural inequalities in school admissions, funding allocation, or certification. We document model-specific disparities and show how open models can be audited for safety. Prompt diversity, few-shot conditioning, and error band analysis further support robust deployment. Institutions can configure open models to avoid overfitting to stylistic markers that disproportionately favor certain writing styles. Ethics: Ethical AI deployment demands transparency, affordability, and user autonomy. Open models support all three. Our study empowers communities to evaluate and choose LLMs based on both performance and values. Open systems allow cultural adaptation, local hosting, and public scrutiny.

We provide a roadmap for moving from proprietary dependence to participatory governance. Our benchmark offers evidence that open and open-source LLMs rival closed models in scoring quality, fairness, and generation fidelity, while offering massive cost savings and ethical advantages. We urge researchers, institutions, and funders to support open evaluation practices and invest in open infrastructure to ensure that LLM capabilities are not limited to those with privileged access. The accessibility divide is a policy and design choice, not a technological inevitability. And it can be bridged through sustained commitment to openness, equity, and RISE-aligned development.

#169: Beyond Model-Centric Interpretations: A Framework for MAcro-eXplainability in Generative AI Systems (MAX AI).

Authors: John Behrens and Alexi Orchard

Abstract: As generative artificial intelligence (GenAI) systems, including large language models (LLMs) and multimodal platforms, become deeply integrated into everyday life, their technical complexity and societal impacts demand new explanatory frameworks. Traditional Explainable AI (XAI) discussions, primarily developed in pre-generative AI contexts, focused predominantly on model-centric interpretations highlighting which input features influenced a specific output using methods like SHAP or LIME, or social concerns in localized, context-specific systems, such as Data Cards, Model Cards (Mitchell et al., 2019) and System Cards(Meta AI, 2022). These approaches emphasized model-centric insights relevant primarily to technical stakeholders and particular system deployments, neglecting broader social interconnections and socio-economic determinants influencing system functionality and use.

Even Ehasan et al's (2021) prescient “Expanding Explainability: Towards Social Transparency in AI systems” only conceptualized social issues within a specific organization. In contrast, this paper proposes a "macro-explainability" framework tailored for the complexities of generative AI, highlighting the interconnected technical, product design, economic, social, and ethical dimensions of AI deployment and use. We believe that In a world where generative AI is directly accessible to all citizens, the requirement for micro-explainability aimed at technical experts is replaced by a broader explainability mandate for the general public.

1. This paper describes the components of a "macro-explainability" framework that expands beyond model mechanics to encompass the entire ecosystem in which AI systems operate. We identify eight key dimensions of macro-explainability Fundamental AI Properties: The inherently data-driven, probabilistic, predictive, and iterative nature of generative AI systems that creates fundamental variability in outputs across both consumer and business applications.
2. Behavioral Dependencies: How system outputs are shaped by input data quality, prompt construction, assigned tasks, and available system tools/constraints.
3. Task Analysis Framework: The structured approaches needed to model and evaluate how AI systems perform across different tasks and contexts.
4. System Behavior Variability: Sources of behavioral inconsistency including training-related variability, task properties, and statistical variation across sample sizes.
5. Product Feature Variability: How diverse implementation choices—training data scope/timing, specialized capabilities (reasoning, coding, multimodal), social behavior variations, tool integration, RLHF processes, guardrails, and user experience features—substantially impact system behavior.
6. Product Support Variability: The range of available documentation, configurable data protection features, and customer support resources that mediate understanding and accountability.
7. Socioeconomic Context: Corporate incentives toward "hype" and user engagement, environmental impacts, labor displacement concerns, and the non-neutrality of technological systems.
8. Human Flourishing: The implications for human intelligence, meaning and work, human dignity, and the dangers of elevating efficiency above human values.

Through a critical assessment of current practices and literature, we discuss how these dimensions collectively provide a more holistic understanding of generative AI behavior than traditional model-centric explanations. Our framework acknowledges that AI systems are not merely technical artifacts but sociotechnical systems embedded in human, organizational, and societal contexts.

By expanding explainability to include these dimensions, we can better address the "black-box" nature of generative AI not just as a computational/statistical algorithm, but as ubiquitous products that require a minimum level of end-to-end understanding for appropriate trust, effective governance, and responsible use. We argue that this shift from micro to macro-explainability is essential for ensuring that increasingly powerful generative AI systems remain accountable and aligned with human values and appropriately used by an informed citizenry.

This framework has directly informed the design and implementation of our course "Generative AI in the Wild" that has been taught for 4 semesters at the University of Notre Dame. The course uses this multidimensional lens to provide students from across the university with both theoretical understanding and hands-on experience across many of these dimensions, moving beyond technical capabilities to critical examination of real-world AI systems. By integrating technical exploration with ethical reflection, the course aims to create an informed citizenry capable of thoughtful engagement with generative AI technologies. Initial student outcomes demonstrate enhanced ability to critically evaluate AI systems across these dimensions, suggesting this macro-explainability framework offers both theoretical value and practical educational utility in preparing society for the responsible deployment and governance of generative AI.

Track: LLMs4All: Large Language Models for Research and Applications in Academic Disciplines

#172: Emergency Algorithms: RISE-compatible Architectures for Times of Crisis

Authors: Daniel Slate

Problem Statement: One domain experiencing increasing uptake of artificial intelligence systems is the field of emergency management, humanitarian assistance, and disaster response. Societies across the world face a grand challenge: the frequent recurrence of natural disasters and human-generated crises, both of increasing number and intensity. The persistent states of crisis and emergency we often find ourselves in are straining our polities' stability, health, and economic integrity. Historically, theories of crisis government entrust emergency powers to entities believed to possess better information and the ability to move faster than a deliberative legislative assembly. Together, both greater insight and speed legitimize delegating emergency authority to an often less-democratic and more opaque decision maker. These same qualities—much in demand during a crisis—will make (and are already making) a compelling case for the use of artificial intelligence systems before, during, and after emergencies, because of AI's proven and expected ability to gather, process, fuse, and present disparate data flows and to do so quickly.

Professionals in the military, law enforcement, public health, humanitarian response, and natural disaster management fields have expressed considerable optimism (some of it well-founded) about AI's advantages and have already begun to deploy such systems when responding to crises, including recent major natural disasters. AI systems are already being used to communicate directly with survivors or those at risk from disasters and to inform emergency responders and relief officials as they decide to whom to distribute life-saving relief. Human decision-makers are thus already relying on AI for the highest stakes choices; it is a short step to using more autonomous and agentic AI systems to adjudicate emergency decisions directly, whether on behalf of public sector entities or adjacent private sector actors in medicine, insurance, and the large networks of volunteer emergency responders.

Surprisingly—given the stakes and the credible forecasts that crisis conditions will increasingly become more the norm than the exception—there has to date been relatively little attention given to designing emergency AI systems in ways that synthesize the ethical principles and legal mandates that govern emergencies. The law has long accepted that the norms that operate during times of crisis can deviate from those that govern during ordinary times. However, those designing emergency AIs need to know that there is more than one way that norms can change during emergencies. What that means is there are important and ultimately unavoidable choices to make about precisely which of several emergency norms an AI should use in times of crisis. At present, however, we are not having the conversation about what emergency ethic (e.g., survivalism or dignitarian realism) we should program and train AIs to consult when making recommendations and decisions.

AI Innovation: This project proposes and comparatively evaluates RISE-consistent AI architectures informed by the different logics of emergency ethics. As an initial design matter, an emergency AI system can either be deployed only during times of crisis or can alternate between ordinary and emergency states. We can expect the latter design to have speed advantages over the former, in that it need not be deployed from a cold start but can hook into existing legal mechanisms that activate and toggle between "normal" and "exception" when governments declare states of emergency.

While in its emergency state, the AI should be able to gather new data, issue instructions to sensor-equipped UAVs to assess the current state and extent of the disaster, process their remote sense data with computer vision algorithms to rapidly identify and prioritize damaged areas and populations, and be capable of writing code, reports, and issuing messages to coordinate official and volunteer responders. This project proposes a fuse-then-forget framework as part of any emergency AI's architecture: the data collected, processed, and inferentially generated while the AI is in its emergency state should be stored in memory separate from ordinary-state memory, so that—solely for the duration of the crisis—it can be treated and processed according to the more lenient emergency exceptions written into current privacy and data protection laws, which otherwise require much more stringent data handling in ordinary times.

Additionally, different emergency AIs can be trained such that their reasoning and chain of thought displays elements consistent with the different emergency ethics: one a purely survivalist AI (hypothesized to be more willing to sacrifice some individuals and rights for the survival of larger groups), another a dignitarian-realist AI (hypothesized to expose reasoning that reveals a willingness to derogate some rights, but never to the extent that human individuals are degraded or discarded). A less autonomous, human-assistive architecture would have similar features: the dignitarian emergency AI assistant would present recommendations to the human emergency management professional alongside reminders that whatever decision the human takes should account for human dignity, perhaps with the additional ability to intervene if a human operator flagrantly ignores such admonitions by alerting others to the dignity-violating override or the ability to suspend or temporarily revoke the human operator's system privileges until other human managers can assess the situation. These are precautions a survivalist AI assistant would not be expected to make.

Translational Evidence: AI is increasingly recognized as inseparable from the future of emergency management and disaster response. Deploying AI systems consistent with the proposed architecture would deliver improved emergency response outcomes with built-in protections for human dignity, which are otherwise often put at risk when governments invoke emergencies. Also, extant legal authorities governing crisis management were set in place before recent technical breakthroughs and do not account for the unique risks and threats that AIs pose, nor adequately specify what kinds of AI-enabled decision-making should be mandated, permitted, or forbidden.

From the legal and policy perspective, AI safety concerns are particularly acute during emergencies. Currently, in both civil law and common law jurisdictions and under international law, emergencies trigger lawful derogations from many of the usual protections and rights individuals enjoy. In the AI context, current emergency laws allow for and enable (explicitly or by implication), the collection, processing, and transfer of data that would normally be protected or inaccessible. Such data is meant to be used only during the time of emergency. However, in an early warning signal, reports have recently surfaced in Australia of AI models being developed on emergency data and then commercialized for use in ordinary, non-exceptional times, in violation of the ethics of consent and the norms of emergency assistance. The fuse-then-forget framework precludes such abuse.

The AI development community should encourage best practices and policies for, and the law should also articulate and codify, a set of requirements for AI developers and deployers that both allows for access to emergency-use-only data during crises (including use through constructive consent) in order to help save lives and also mandates secluding that data from the ordinary, ongoing training processes of the models, so that it is not commodified for use in ordinary times. Such regulation will help align AI systems with dignity-respecting emergency norms.

RISE Dimensions: Current early efforts to design and deploy AI to help before, during, and after emergencies offer us a rare opportunity to re-open choices once thought settled by legal and political theorists. In particular, the dominant emergency ethic is a norm of mere survivalism, often operating under the name of the law of necessity (or the idea that "necessity knows no law"), which instructs that, during emergencies, nearly all values must fall before the survival of the main part of the political community. An alternative emergency ethic exists, however, that combines the realism of survival with an inviolable requirement to respect the worth and dignity of every human being. While the currently dominant emergency ethic has its roots in the office of the dictator of ancient Rome, the alternative ethic has equally ancient origins in the Talmud and has begun to reappear in major international law treaties (e.g., the ICCPR) that prioritize human dignity even in times of exigency.

Additionally, there is a phenomenon well known to emergency managers and disaster response professionals, particularly those operating in locales with authoritarian governments. Many vulnerable populations do not want to be found or noticed by the authorities, but it is often the government that insists on taking the lead during crises. As a result, vulnerable populations including refugees, dissenters, minorities, and the politically disenfranchised in authoritarian countries often seek to evade the attention of emergency responders, out of fear that their existence will then become known, exposing them to persecution, repression, or worse. The fuse-then-forget emergency AI architecture proposed here offers a solution to this problem: if it is the autonomous emergency-mode AI that is coordinating emergency response efforts and parsing data from ground and overhead sensors, rather than the authoritarian government's personnel, then the latter never need to see the data and inferences that would reveal (and expose to harm) the existence and locations of vulnerable communities that are trying to evade the gaze of their authoritarian state.

Vulnerable populations could thus receive humanitarian assistance safely in ways not previously possible. This project's approach thus suggests that AI-powered emergency response could be more inclusive than traditional purely human response efforts. The AI safety community has given significant attention, and rightly so, to how powerful AI systems might threaten or damage our societies. This project attends to the subset of AI systems that can help protect and repair human communities in the wake of catastrophes, offering a set of proposals to direct the development and governance of such systems in inclusive ways that improve, rather than worsen, human safety and dignity.