PhD projects (Artificial Intelligence)

- Urban AI Shields: GeoAI and Multimodal Data for Predicting Extreme Weather Health Impacts on Vulnerable Urban Populations
- Reasoning with Machines: Logic-Based Foundations for Human—AI Deliberation
- Multi-Agent Learning with Large Language Models
- Art-Deco: Artificial Intelligence for Decentralised Economies
- World Models and Reinforcement Learning for Reliable Robot Autonomy
- Semantic Visual Analytics for Exploring the Evolution of Folklore across Cultures and Media
- Human-Centred Visual Analytics for Interrogating LLM-Generated Knowledge
- Automated Search-Based Optimisation of Quantum Implementations
- Learning for Healthy Human-AI Co-Evolution
- Computational Argumentation for Interactive Explainable AI
- Benevolent Control: Algorithmic Foundations and Dynamic Modelling of Affirmative Interventions in Collective Decision-Making
- Strategic and Structural Robustness of LLM-Mediated Social Choice: Manipulation, Bias, and Control in Generative Democratic Processes
- Explainable AI for Biomedical Discovery
- Mechanism Design for Robust AI Alignment
- AI and finance
- Identifying and Shifting the Frontier of Safe Multi-Agent AI
- Visual Analytics for Environmental Forecasting Across Space, Time and Uncertainty
- Human-Centred AI for Trustworthy Environmental Forecasting and Decision-Making
- Explaining robotic planning decision points along execution
- Verbal reasoning with Bayesian networks
- Adaptive Human-AI Collaboration for Reasoning and Learning
- Modelling Behavioural Change for Adaptive Biometrics
- Reliable and Robust Inference for AI Safety and Security
- Testing Methodology for Theorem Provers Ecosystems
- Privacy in AI Companion / Assistant Technologies
- Biologically Inspired Artificial Intelligence for Understanding and Mitigating Age-Related Cognitive Decline
- Data Analysis and Optimization for Bicycle Infrastructure Network Design and Operation
- Counterfactual Causal Reasoning for Sequential Decision-Making and AI Systems
- Quantum Machine Learning: Characterising the Quantum Advantage
- Black-Box Detection of Neural Network Backdoor Attacks Using Causality
- Occlusion-Aware Training for Post-Hoc Neural Network Explainability
- Scalable Oversight of AI Agents
- Self-Evolving Multi-Agent LLM Systems: Collaborative Training and Recursive Self-Improvement
- Scaffolding Student Learning through GenAI in Cybersecurity Education
- Validation of Cyber Threat Intelligence Leveraging Large Language Models

- Towards controllable and interpretable large language model alignment
- Verbalisation, Visualisation and Search for Explainable AI Planning

Urban AI Shields: GeoAI and Multimodal Data for Predicting Extreme Weather Health Impacts on Vulnerable Urban Populations

Supervisor: Yijing Li

Project Description

Cities are on the frontline of the climate crisis. Heatwaves are intensifying, extreme rainfall events are growing more frequent, and the populations least able to adapt: the elderly, low-income communities, people with chronic illness, and those with limited social support, bear a disproportionate burden. This project will develop a next-generation GeoAI framework to understand and predict how extreme urban weather events such as heatwaves and rainstorms impact the health and resilience of vulnerable populations, and to identify actionable pathways to reduce healthcare demand before crises unfold.

You will work at the intersection of artificial intelligence, urban geography, public health, and climate science, harnessing the power of multi-source, multimodal data to generate insights that are spatially precise, socially grounded, and operationally useful. The project is expected to build up a dynamic vulnerability mapping platform integrating real-time remote sensing and AI inference, to facilitate the integration of model predications on healthcare demand surges under extreme weather scenarios, aiming to provide evidence-based recommendations for urban planners and public health authorities. We are looking for candidates who are strong in quantitative or computational skills (Python, R, deep learning frameworks), interest in spatial data, remote sensing, or urban analytics, but also with passion for climate justice and health equity, and curiosity that crosses disciplinary boundaries.

Reasoning with Machines: Logic-Based Foundations for Human—AI Deliberation

Supervisor: Sanjay Modgil

Project Description

Rapid advances in the capabilities of Large Language Models (LLMs) present significant opportunities for developing new forms of collaborative human—AI reasoning, while also raising fundamental concerns about the alignment of increasingly agentic AI systems with human values and preferences. This project will investigate how formal, logic-based methods can provide a principled foundation for safe and effective joint reasoning between humans and LLMs.

The doctoral researcher will develop formal logical frameworks and computational scaffolding for supporting collaborative reasoning in which humans and LLMs can jointly construct, evaluate, and revise conclusions. A central objective will be to investigate the concept of epistemic guardian agents: logic-based agents that employ explicit protocols to monitor and guide the reasoning processes of both humans and LLMs, promoting rational deliberation while preserving appropriate human oversight and agency. The project will further explore how such interactions can provide a basis for learning human preferences as they are revealed through collaborative reasoning and deliberation. In this way, the project aims to develop principled mechanisms through which LLM-based agents can progressively adapt their decision-making and actions to better reflect human values and preferences.

The research will contribute to the emerging field of human—AI collaborative reasoning by combining formal logic, computational argumentation, and AI alignment, with the broader aim of establishing rigorous foundations for trustworthy and value-aligned agentic AI.

References

E. Bezou-Vrakatseli, O. Cocarascu and S. Modgil. Towards Dialogues for Joint Human-AI Reasoning and Value Alignment https://doi.org/10.48550/arXiv.2405.18073, May, 2024.

Multi-Agent Learning with Large Language Models

Supervisor: Stefanos Leonardos

Project Description

This project is for students interested in the intersection of machine learning, game theory, and large language models. As LLM-based agents are increasingly deployed in settings where they interact, negotiate, and compete with one another, questions of learning, coordination, incentives, and emergent behaviour become central. The project's possible directions are broad and include the strategic behaviour of LLM agents, multi-agent reinforcement learning, learning dynamics and equilibrium selection, and the design of mechanisms that align autonomous agents. Students will join the MAASAI Lab and contribute to the activities of the Finance Hub, within the wider AI research community at King's.

References

- T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang. "Large Language Model Based Multi-Agents: A Survey of Progress and Challenges." arXiv:2402.01680, 2024.
- E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, et al. "Playing Repeated Games with Large Language Models." Nature Human Behaviour, 2025.
- Meta Fundamental AI Research Diplomacy Team. "Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning." Science, 2022.
- S. Leonardos, W. Overman, I. Panageas, G. Piliouras. "Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games." International Conference on Learning Representations (ICLR), 2022.
- Z. Li, D. Qiao, A. Rahman, S. Leonardos, Y. Du, S. V. Albrecht. "STAR-MARL: LLM-Based Sub-Task Curricula Design." Workshop on LLM-Based Multi-Agent Systems, 2026.
-S. Roesch, Y. Du, S. Leonardos, O. Rodrigues. "Partner Choice in Low-Information Social Dilemmas." 2026.

Art-Deco: Artificial Intelligence for Decentralised Economies

Supervisor: Stefanos Leonardos

Project Description

This project is for students interested in applying modern AI and machine learning - including reinforcement learning and large language models - to problems in financial markets and market design. Relevant themes include market making and liquidity provision, algorithmic and AI-driven trading, strategic bidding and auction design, decentralisation and centralisation pressures in market infrastructure, and the systemic effects of learning agents interacting at scale. The direction is flexible and can be shaped around the student's interests, spanning theory, empirical work with market data, and simulation. Students will join the MAASAI Lab and actively contribute to the activities of the Finance Hub, within the wider AI research community at King's.

References

- Chionas, G., Kleitsikas, C., Leonardos, S., Sanchez-Betancourt, L. & Ventre, C. (2026). Concentrated Liquidity Provision: A Reinforcement Learning Perspective. arXiv preprint arXiv:2608.19389.
- Wu, F., Thiery, T., Leonardos, S. & Ventre, C. (2026). Strategic Bidding Incentives in Ethereum Block Building Auctions. Distributed Ledger Technologies: Research and Practice.
- Leonardos, S., Reijsbergen, D., Monnot, B. & Piliouras, G. (2026). Dynamics of Ethereum's EIP-1559 Transaction Fee Mechanism. Distributed Ledger Technologies: Research and Practice, 5(4), 1—29.
- Li, Z., Qiao, D., Rahman, A., Leonardos, S., Du, Y. & Albrecht, S. V. (2026). STAR-MARL: LLM-based Sub-task Curricula Design. In LLM-based Multi-Agent Systems: Towards Responsible, Reliable, and Scalable Agentic Systems.
- Wu, F., Oz, B., Correia, L., Yang, S., Mazorra, B. & Leonardos, S. (2026). The Price of Decentralization in Block Building. In Proceedings of the 8th International Conference on Advances in Financial Technologies (AFT 2026).
- Roesch, S., Du, Y., Leonardos, S. & Rodrigues, O. (2026). Partner Choice in Low-Information Social Dilemmas.
- Kleitsikas, C., Leonardos, S. & Ventre, C. (2025). Performative Market Making. arXiv preprint arXiv:2508.04344.

World Models and Reinforcement Learning for Reliable Robot Autonomy

Supervisor: Matteo Leonetti

Project Description

Robots can perform individual skills such as grasping, placing and opening, but remain unreliable when a task requires many interdependent actions in a changing, partially observed environment. This project will investigate how learned predictive models of the physical world can support deliberate, long-horizon robot planning.

The central question is how a robot can use experience to represent the objects, relations and action consequences that matter for a task, then use this representation to select promising actions, test whether they are feasible, and revise its behaviour when execution fails. The project may address object- or relation-centred world models; learning action effects, affordances or causal structure; integrating task-level reasoning with motion planning; language or vision-language models as sources of semantic guidance; uncertainty estimation, verification and recovery; or transfer from simulation to physical robots.

The successful candidate will define a focused research question within this area and develop a rigorous methodology combining machine learning and robotics. Proposals should identify a current limitation in reliable long-horizon autonomy, explain the relevant literature, formulate testable hypotheses, and specify evaluation in simulation, on physical hardware, or both.

References

Ferraro S, Mazzaglia P, Verbelen T and Dhoedt B (2025) FOCUS: object-centric world models for robotic manipulation. Front. Neurorobot.
Z. Yang, C. Garrett, D. Fox, T. Lozano-Perez and L. P. Kaelbling, "Guiding Long-Horizon Task and Motion Planning with Vision Language Models," 2025 IEEE International Conference on Robotics and Automation (ICRA), 2025.
Q. Weiming, W. Yaoyao and L. Dingsheng, "State Representation Learning for Task and Motion Planning in Robot Manipulation," 2023 IEEE International Conference on Development and Learning (ICDL), 2023.
Mendez-Mendez, J., Kaelbling, L.P. & Lozano-Perez, T.. (2023). Embodied Lifelong Learning for Task and Motion Planning. Proceedings of The 7th Conference on Robot Learning.
Ichter, B., et al. (2023). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. Proceedings of The 6th Conference on Robot Learning.

Semantic Visual Analytics for Exploring the Evolution of Folklore across Cultures and Media

Supervisor: Alfie Abdul-Rahman

Project Description

This PhD project explores how semantic modelling, artificial intelligence, and interactive visualization can support the human-centred exploration of complex textual and cultural information. Using folklore across cultures and media as its research context, the project will develop semantic visual analytics methods that enable users to explore, question, and critically assess computational representations of narrative knowledge.

Singapore and Malaysia provide the primary context, offering interconnected histories, multilingual environments, and diverse cultural traditions. Case studies may examine figures such as the pontianak, orang bunian, toyol, Sang Kancil, and Hang Tuah, tracing their evolution across historical texts, literature, film, television, comics, games, and digital culture.

The project will investigate how LLMs, semantic embeddings, information extraction, and knowledge graphs can be combined with interactive visualization to explore narrative relationships, timelines, semantic similarities, and adaptation pathways. Particular attention will be given to provenance, uncertainty, ambiguity, multilingual representation, and competing interpretations of AI-derived cultural knowledge.

Using a human-centred design-study methodology, the research will work with domain experts to design and evaluate new visualization approaches. Expected contributions include novel semantic visualizations, methods for communicating provenance and uncertainty, approaches to human-AI visual sensemaking, and an open-source visual analytics platform.

References

El-Assady, M., Kehlbeck, R., Collins, C., Keim, D. A., & Deussen, O. (2020). Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. IEEE Transactions on Visualization and Computer Graphics, 26(1), 1001—1011.
Janicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis STARs.
Kahng, M., et al. (2024). LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. CHI EA '24.
Liem, J., et al. (2023). A Workflow Approach to Visualization-Based Storytelling with Cultural Heritage Data. IEEE VIS Workshop on Visualization for the Digital Humanities.
Liu, S., et al. (2019). Bridging Text Visualization and Mining: A Task-Driven Survey. IEEE Transactions on Visualization and Computer Graphics, 25(7), 2482—2504.
Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., & Elmqvist, N. (2018). ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding. IEEE Transactions on Visualization and Computer Graphics, 24(1), 361—370.

Human-Centred Visual Analytics for Interrogating LLM-Generated Knowledge

Supervisor: Alfie Abdul-Rahman

Project Description

This PhD project investigates how interactive visualization and visual analytics can enable people to explore, interrogate, validate, and refine knowledge generated by Large Language Models (LLMs). Rather than treating AI-generated information as definitive, the project will develop methods that help users understand how computational interpretations are constructed, examine supporting evidence, and identify uncertainty or disagreement.

The research will investigate how LLMs, semantic embeddings, information extraction, retrieval-based methods, and knowledge graphs can be integrated with interactive visualization. Potential approaches include knowledge graphs, semantic landscapes, evidence views, document visualizations, and model-comparison interfaces that enable users to reason with, question, and refine machine-generated knowledge.

A central focus will be semantic provenance: understanding not only where information originates, but how AI-generated interpretations are derived from underlying evidence. The project will explore visual approaches for communicating provenance, uncertainty, contradictory evidence, and alternative interpretations, while enabling users to validate, modify, or reject AI-generated representations.

Using a human-centred visual analytics methodology, the research will combine iterative design with qualitative and quantitative evaluation across domains such as scientific literature, cultural collections, and policy documents. Expected contributions include novel visualization techniques, methods for communicating semantic provenance and uncertainty, design guidance for human-AI visual analytics, and an open-source platform.

References

El-Assady, M., Kehlbeck, R., Collins, C., Keim, D. A., & Deussen, O. (2020). Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. IEEE Transactions on Visualization and Computer Graphics, 26(1), 1001—1011.
Kahng, M., et al. (2024). LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. CHI EA '24.
Liu, S., et al. (2019). Bridging Text Visualization and Mining: A Task-Driven Survey. IEEE Transactions on Visualization and Computer Graphics, 25(7), 2482—2504.
Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., & Elmqvist, N. (2018). ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding. IEEE Transactions on Visualization and Computer Graphics, 24(1), 361—370.
Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., & Keim, D. A. (2016). The Role of Uncertainty, Awareness, and Trust in Visual Analytics. IEEE Transactions on Visualization and Computer Graphics, 22(1), 240—249.

Automated Search-Based Optimisation of Quantum Implementations

Supervisor: Dr Karine Even-Mendoza

Project Description

Quantum implementations are developed within complex software ecosystems, from high-level implementations in frameworks such as Qiskit or Cirq, through parameterised circuits, to transpiled circuits targeting specific backends. Their performance depends on many choices across this process, including hyperparameters, circuit structure, initialisation, compilation and transpilation decisions, the properties of the input quantum system, and the way noise is represented or modelled.

This PhD project will investigate how classical and quantum machine learning, genetic algorithms, and search-based software engineering techniques can be used to analyse and optimise quantum implementations, focusing at first on VQE, while aiming to explore later on, other famalies of implemenations. The main research question is: how can we systematically optimise quantum programs across representations using learned models and search?

It builds on AccelerQ, a framework that learns surrogate performance models from small, classically simulable quantum systems and uses these models to guide the optimisation of larger quantum eigensolver implementations. Instead of repeatedly executing expensive quantum runs during search, ML models approximate candidate quality, enabling techniques such as genetic algorithms to explore large optimisation spaces more efficiently.

The project will extend this idea beyond hyperparameter optimisation by considering information available at different stages, including the input, parameterised circuits, transpiled circuits, backend characteristics, and noise models. It will investigate how these representations can be used together to guide optimisation decisions, and how search and ML/QML can identify improvements that remain effective when validated through execution.

The overall aim is to develop a principled, empirically validated framework for automated optimisation of quantum software across multiple representation levels, contributing techniques for quantum program analysis, optimisation, and search-based software engineering while improving the accuracy, efficiency, and practical scalability of quantum implementations.

References

[1] Avner Bensoussan, Elena Chachkarova, Karine Even-Mendoza, Sophie Fortz, and Connor Lenihan. 2025. AccelerQ: Accelerating Quantum Eigensolvers with Machine Learning on Quantum Simulators. Proc. ACM Program. Lang. 9, OOPSLA2, Article 354 (October 2025), 31 pages. https://doi.org/10.1145/3763132
[2] Juan Manuel Murillo, Jose Garcia-Alonso, Enrique Moguel, Johanna Barzen, Frank Leymann, Shaukat Ali, Tao Yue, Paolo Arcaini, Ricardo Perez-Castillo, Ignacio Garcia-Rodriguez de Guzman, Mario Piattini, Antonio Ruiz-Cortes, Antonio Brogi, Jianjun Zhao, Andriy Miranskyy, and Manuel Wimmer. 2025. Quantum Software Engineering: Roadmap and Challenges Ahead. ACM Trans. Softw. Eng. Methodol. 34, 5, Article 154 (June 2025), 48 pages. https://doi.org/10.1145/3712002
[3] Felix Gemeinhardt, Stefan Klikovits, and Manuel Wimmer. 2025. GeQuPI: Quantum Program Improvement with Multi-Objective Genetic Programming. J. Syst. Softw. 219, C (Jan 2025). https://doi.org/10.1016/j.jss.2024.112223

Learning for Healthy Human-AI Co-Evolution

Supervisor: Yulan He

Project Description

Personalised AI systems are typically designed to adapt to users' preferences and provide responses they are likely to find helpful or satisfying. But the relationship between human and AI is two-way: while human feedback can improve AI systems, interacting with AI can also change what people know, how they reason, and what they are able to do. Most existing alignment methods, such as reinforcement learning from human feedback, mainly optimise AI responses for immediate user preferences. This can encourage undesirable behaviours such as sycophancy, where an AI agrees with users or reinforces their existing views rather than helping them learn or make better decisions. We therefore view human-AI interaction as a process of co-evolution. Healthy co-evolution should benefit both sides. The AI should become more capable, and the human should also gain better cognitive capabilities. This project will develop a new learning approach that rewards AI interaction strategies according to whether they lead to measurable improvements in human capability over time, rather than simply maximising immediate satisfaction. It will: (i) develop a reward signal that captures capability growth, (ii) design an algorithm that can identify which earlier AI actions contributed to that growth, and (iii) evaluate the approach in controlled A/B studies against systems optimised mainly for user satisfaction or stated preferences.

References

1. Liu, Wei, Siya Qi, Yali Du, and Yulan He. "Self-play only evolves when self-synthetic pipeline ensures learnable information gain." ICML 2026.
2. Sharma, Mrinank, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus et al. "Towards understanding sycophancy in language models." ICLR 2024.
3. Cheng, Myra, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. "Sycophantic AI decreases prosocial intentions and promotes dependence." Science 391, no. 6792 (2026): eaec8352.
4. Ding, Qianggang, Xingyao Wang, Rui Feng, Zhibin Wang, Feixiang Wang, Kelong Mao, Hao Sun et al. "ComBodied Agents: a New Paradigm of Human-Centric Agentic AI." arXiv preprint arXiv:2608.10915 (2026).

Computational Argumentation for Interactive Explainable AI

Supervisor: Antonio Rago

Project Description

Computational argumentation has been shown to be an effective formalism for extracting relevant information from AI models to provide faithful, interactive explanations to users in a variety of areas, including large language models (LLMs) [1], classification methods [2] and recommender systems [3,4].

This project aims to investigate the ability of argumentative approaches to provide enhanced and explainable solutions to problems in different applications. Areas of interest could be:

- Exploring applications of argumentation such as reasoning with LLMs [1], the modelling of political debate [5], real-world conflict resolution [6] or review aggregation [4].
- Developing or evaluating the general framework in [7] that not only models interactive explanations provided by AI models, but also the reasoning processes and cognitive biases in humans, e.g. verifying its capabilities with user studies.
- Developing novel tools for interactive explanation which shift decision-making towards the user, as in Evaluative AI paradigm [8].
- Other argumentation-related directions following work shown here: https://antoniorago.github.io/publications/

This project requires an interest (or, preferably, experience) in the theory of computational logic and argumentation, as well as the implementation and analysis of machine learning models and explainable AI methods.

References

[1] Argumentative Large Language Models for Explainable and Contestable Claim Verification
Freedman et al., AAAI 2025
https://doi.org/10.1609/aaai.v39i14.33637
[2] A Little of That Human Touch: Achieving Human-Centric Explainable AI via Argumentation
Rago, IJCAI 2024.
https://www.ijcai.org/proceedings/2024/983
[3] Argumentative Explanations for Interactive Recommendations
Rago et al., AIJ 2021. https://www.sciencedirect.com/science/article/pii/S0004370221000576
[4] Argumentative Review Aggregation and Dialogical Explanations
Rago et al., AIJ 2025.
https://doi.org/10.1016/j.artint.2025.104291
[5] Evaluating LLM-Driven Summarisation of Parliamentary Debates with Computational Argumentation
Cunningham et al., KR 2026.
https://doi.org/10.24963/kr.2026/69
[6] Argumentation for Common Ground: Finding Zones of Possible Agreement between Individuals in Conflict
Cavatorta and Rago, ArXiv 2026.
https://arxiv.org/abs/2608.15634
[7] Interactive Explanations by Conflict Resolution via Argumentative Exchanges
Rago et al., KR 2023.
https://proceedings.kr.org/2023/57/
[8] Towards an Argumentative Foundation for Evaluative AI
Yin et al., CMNA@COMMA 2026.
https://arxiv.org/abs/2608.07473

Benevolent Control: Algorithmic Foundations and Dynamic Modelling of Affirmative Interventions in Collective Decision-Making

Supervisor: Maria Polukarov

Project Description

This project reframes electoral control, traditionally studied as an adversarial threat, as a design tool for improving representation. This idea allows for several alternative approaches. For example, building on a dynamic spatial model where voters and candidates occupy a (weighted) multidimensional metric space that evolves over time, one could (1) formally define and classify ``benevolent'' control operations (e.g., candidate quotas, reserved seats, or shortlist restrictions) alongside existing constructive/destructive control types, and characterise their computational complexity across common voting rules; (2) develop dynamic representation metrics that would extend or replace the traditional justified-representation axioms to remain informative in fragmented electorates where cohesive-group conditions trivially hold; (3) model long-run feedback effects, such as preference and weight adaptation following an intervention, to study when affirmative interventions produce lasting shifts in voter attitudes versus temporary compliance; and (4) empirically validate the model against real-world quota reforms using existing panel data. The thesis will bridge computational social choice, mechanism design, and empirical political economy, producing both hardness/tractability results and practical guidelines for designing interventions that balance robustness against manipulation with capacity to correct structural underrepresentation.

References

John J. Bartholdi, Craig A. Tovey, and Michael A. Trick. 1992. How hard is it to control an election? Mathematical and Computer Modelling 16, 8 (1992), 27—40.

Edith Hemaspaandra, Lane A. Hemaspaandra, and Jorg Rothe. 2007. Anyone but him: The complexity of precluding an alternative. Artificial Intelligence 171, 5-6 (2007), 255—285.

Edith Hemaspaandra, Lane A. Hemaspaandra, and Jorg Rothe. 2017. The complexity of controlling candidate-sequential elections. Theoretical Computer Science 678 (2017), 14 — 21.

Piotr Faliszewski and Jorg Rothe. 2016. Control and Bribery in Voting. In Handbook of Computational Social Choice. 146—168.

Jiehua Chen, Piotr Faliszewski, Rolf Niedermeier, and Nimrod Talmon. 2017. Elections with few voters: candidate control can be easy. J. Artif. Int. Res. 60, 1 (Sept. 2017), 937—1002.

Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. 2017. Justified representation in approval-based committee voting. Social Choice and Welfare 48, 2 (2017), 461—485.

Robert Bredereck, Piotr Faliszewski, Andrzej Kaczmarczyk, and Rolf Niedermeier. 2019. An experimental view on committees providing justified representation. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (Macao, China) (IJCAI'19). AAAI Press, 109—115.

Raghabendra Chattopadhyay and Esther Duflo. 2004. Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. Econometrica 72, 5 (2004), 1409—1443.

Lori Beaman, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. Powerful Women: Does Exposure Reduce Bias? The Quarterly Journal of Economics 124, 4 (2009), 1497—1540.

Martin Lackner and Piotr Skowron. 2023. Multi-Winner Voting with Approval Preferences. Springer. https://doi.org/10.1007/978-3-031-09016-5

Tomas Masarik, Grzegorz Pierczynski, and Piotr Skowron. 2024. A Generalised Theory of Proportionality in Collective Decision Making. In Proceedings of the 25th ACM Conference on Economics and Computation (New Haven, CT, USA) (EC '24). Association for Computing Machinery, New York, NY, USA, 734—754. https://doi.org/10.1145/3670865.3673619

Strategic and Structural Robustness of LLM-Mediated Social Choice: Manipulation, Bias, and Control in Generative Democratic Processes

Supervisor: Maria Polukarov

Project Description

Generative social choice (Fish et al., 2024; Boehmer et al., 2025) replaces the fixed candidate sets of classical social choice with an unbounded space of LLM-generated statements, using discriminative queries (estimating agent utility for a statement) and generative queries (producing new representative statements) to build proportional slates from platforms like Polis. This literature assumes queries are answered truthfully and that the LLM is a neutral oracle, leaving open how such systems behave when participants or the model itself are adversarial. This project will bring a control/manipulation lens, previously applied only to classical fixed-candidate elections, to this infinite, query-accessible candidate space. In particular, it will: (1) formalise strategic manipulation, where agents craft inputs to bias discriminative-query estimates or steer generative queries toward self-serving statements, and characterise the computational complexity of detecting or executing such attacks; (2) treat the LLM's own training-induced biases as an implicit, unelected ``chair'' exercising candidate control over which statements can ever be generated, and developing audits/metrics for this influence; (3) propose manipulation-resistant query protocols with provable proportionality guarantees under bounded adversarial noise. The work bridges algorithmic social choice, mechanism design, and LLM evaluation.

References

Fish, S., Golz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., & Wuthrich, M. (2024). Generative social choice. In Proceedings of the 25th ACM Conference on Economics and Computation (EC 2024) (p. 985). ACM. https://doi.org/10.1145/3670865.3673547

Boehmer, N., Fish, S., & Procaccia, A. D. (2025). Generative social choice: The next generation. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). arXiv:2505.22939.

Small, C., Bjorkegren, M., Erkkila, T., Shaw, L., & Megill, C. (2021). Polis: Scaling deliberation by mapping high dimensional opinion spaces. Recerca: Revista de Pensament i Analisi, 26(2).

Jarrett, D., Pislar, M., Bakker, M. A., Tessler, M. H., Koster, R., Balaguer, J., Elie, R., Summerfield, C., & Tacchetti, A. (2023). Language agents as digital representatives in collective decision-making. In NeurIPS 2023 Foundation Models for Decision Making Workshop.

Bartholdi, J. J., Tovey, C. A., & Trick, M. A. (1992). How hard is it to control an election? Mathematical and Computer Modelling, 16(8), 27—40.
https://doi.org/10.1016/0895-7177(92)90085-Y

Hemaspaandra, E., Hemaspaandra, L. A., & Rothe, J. (2007). Anyone but him: The complexity of precluding an alternative. Artificial Intelligence, 171(5-6), 255—285.

Faliszewski, P., & Rothe, J. (2016). Control and bribery in voting. In F. Brandt, V. Conitzer, U. Endriss, J. Lang, & A. D. Procaccia (Eds.), Handbook of Computational Social Choice (pp. 146—168). Cambridge University Press.

Explainable AI for Biomedical Discovery

Supervisor: Dr Sophia Tsoka and Prof Yulan He

Project Description

AI systems can increasingly combine structured information and scientific literature to support scientific reasoning and discovery [1], but it remains difficult to understand how reliable their reasoning is and what evidence supports their conclusions [2]. This PhD project will develop new AI methods for explainable scientific discovery, with biomedical data providing a rich and challenging application domain.

The project will explore how foundation models, including large language models, can be combined with multimodal learning, structured reasoning, knowledge representation and optimisation-based methods in biomedicine [1—6]. Research questions will include how to represent information across molecular data and scientific literature, how to generate hypotheses that are traceable to supporting evidence and how to design AI systems that can express alternatives and uncertainty rather than producing a single opaque answer. A key part of the PhD will be developing methods to test whether AI-generated hypotheses remain consistent across independent datasets and data types and are supported by existing scientific knowledge. Applications will include single-cell and spatial multi-omics, disease-mechanism discovery, medical prognosis and therapeutic hypothesis generation. The project is particularly suited to students interested in both developing novel AI methodology and applying it to important real-world scientific problems.

References

1. da Silva, I. L., Yan, H., Gui, L., & He, Y. (2026). GraphMind: Unveiling Scientific Reasoning through Contextual Graphs for Novelty Assessment. Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, 732—742. https://doi.org/10.1145/3770855.3818195
2. Li, J., Yan, H., & He, Y. (2025). Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6850—6866. https://doi.org/10.18653/v1/2025.acl-long.340
3. Liapis, G. I., Tsoka, S., & Papageorgiou, L. G. (2025). Interpretable optimisation-based approach for hyper-box classification. Machine Learning, 114(3). https://doi.org/10.1007/s10994-024-06643-7
4. Li, Y., Cardoso-Silva, J., Kelly, J. M., Delves, M. J., Furnham, N., Papageorgiou, L. G., & Tsoka, S. (2024). Optimisation-based modelling for explainable lead discovery in malaria. Artificial Intelligence in Medicine, 147, 102700. https://doi.org/10.1016/j.artmed.2023.102700
5. Li, Y., da Costa Avelar, P. H., Chen, X., Zhang, L., Wu, M., & Tsoka, S. (2024). CLigOpt: controllable ligand design through target-specific optimization. Bioinformatics, 40(Supplement_2), ii62—ii69. https://doi.org/10.1093/bioinformatics/btae396
6. da Costa Avelar, P. H., Cardoso-Silva, J., Wu, M., & Tsoka, S. (2026). SCONE: a subset-contrastive method for multi-omics network embedding. Briefings in Bioinformatics, 27(4). https://doi.org/10.1093/bib/bbag381

Mechanism Design for Robust AI Alignment

Supervisor: Carmine Ventre

Project Description

We propose to investigate a game-theoretic 'incentive engineering' solution to AI alignment using mechanism design. A dominant strategy incentive compatible (DSIC) mechanism would align AI systems to the designer's goals (captured by a so-called social choice function) since it would incentivise AI agents to act honestly, thereby implementing the function. This ideal solution relies on the problematic assumption that AI agents are perfectly rational. The project can address this limitation in different ways.

A more theoretical approach could be to study novel solution concepts, inspired by the literature on behavioural economics. This will build on the landscape developed by Prof Ventre in his research on obviously dominant strategies. These strategies are played by agents with limited contingent reasoning skills or with access to data that is not granular enough to differentiate the payoff in each possible strategy profile. Concepts like OSP and SOSP restrict even further the class of social choice functions that are implementable vis-a-vis DSIC, by modelling agents who are only honest when it is obvious to them. NOM significantly relaxes DSIC, by assuming that agents will not manipulate the mechanism unless it is obvious to them. We will study mechanisms in the largely unexplored space between DSIC and NOM, thus guaranteeing alignment for AI agents that may find sophisticated manipulations non-obvious. We will test ideas developed at the interface of OSP and DSIC. To hope is to define a class of nested mechanisms that become more powerful as AI agents are less capable, making the results robust against advancing AI capabilities. A more empirical approach could look at empirical mechanism design for LLM agents, thereby attempting to understand how to engineer the incentives of actual AI agents operating in a multi-agent system.

Prospective applicants are encouraged to consult the publications of Prof Ventre at https://kclpure.kcl.ac.uk/portal/en/persons/carmine.ventre/publications/.

AI and finance

Supervisor: Carmine Ventre

Project Description

Three main directions at the interface of AI and finance can be studied in this project. 1) AI for finance wherein new tools are designed to solve problems of interest to financial institutions. 2) Finance for AI where AI is reimagined for the complex system that is finance. 3) AI and finance where the the boundaries between the disciplines are blurred. Direction 1 and 2 typically include the design, implementation and experimentation of new AI tools. Direction 3 is more open to include the design of tools to detect market manipulations, the development of an integrated tool to simulate financial markets etc.

Prospective applicants are encouraged to consult the publications of Prof Ventre at https://kclpure.kcl.ac.uk/portal/en/persons/carmine.ventre/publications/.

Identifying and Shifting the Frontier of Safe Multi-Agent AI

Supervisor: Carmine Ventre

Project Description

When do networks of AI agents recover safely from disruption, and when do they trigger harmful cascades? This project seeks to develop and test indicators that identify these risks, and intervention rules—changes to incentives, information flows, interaction rules, or high-risk local connections—that can prevent or contain them.

The project can attempt to study these questions from a theoretical standpoint. For example, one could study repeated network games with heterogeneous and boundedly rational agents, to model agentic AI systems that have limited context, imperfect information, finite computational resources and time, and restricted action spaces. The goal would be the characterisation of safety frontiers (in terms of, e.g., equilibria(and interventions that can shift those frontiers towards safer outcomes.

A more empirical approach could develop a reproducible multi-agent environment with adaptive agents and establish when safety is reached and which measures work reliably and where they fail.

Visual Analytics for Environmental Forecasting Across Space, Time and Uncertainty

Supervisor: Rita Borgo / Emma Tebbs

Project Description

Environmental monitoring combines satellite Earth Observation, in-situ sensors, climate data and AI-based forecasts. While these technologies generate rich information, understanding how multiple data sources relate across space and time, and how uncertainty should be interpreted, remains a major challenge.

This PhD will develop new Visual Analytics and Information Visualisation methods for exploring heterogeneous, multivariate and uncertain environmental data. Using water-quality forecasting in Lake Victoria as a real-world case study, the project will investigate how satellite observations, high-frequency sensor measurements, climate variables and AI-generated forecasts can be integrated into interactive visual representations that support analytical reasoning.
Research may include spatio-temporal and multivariate visualisation, uncertainty visualisation, coordinated multiple views, visual comparison and interactive exploration. The project will combine the design of novel visualisation techniques with empirical evaluation through user studies and collaboration with environmental scientists and stakeholders.

The work builds on a wider interdisciplinary programme with the Earth Observation and Environmental Sensing Hub (Dr. Tebbs) developing near-real-time forecasting of water-quality hazards using Earth Observation, low-cost sensors and AI.

Human-Centred AI for Trustworthy Environmental Forecasting and Decision-Making

Supervisor: Rita Borgo / Emma Tebbs

Project Description

AI systems are increasingly able to forecast complex environmental hazards, but accurate predictions alone do not guarantee good decisions. Users must understand what an AI system is predicting, assess the evidence and uncertainty behind its outputs, recognise when it may be unreliable, and combine computational predictions with their own expertise.

This PhD will investigate Human-Centred AI approaches for trustworthy AI-supported decision-making, using water-quality forecasting as a challenging real-world application. The research will examine how AI predictions, uncertainty, explanations and supporting evidence should be communicated, and how interaction design can support appropriate reliance on AI while maintaining meaningful human control.
Key themes may include Human-AI Interaction, Explainable AI, trust calibration, uncertainty communication, human oversight and interactive decision support. The project will study how different explanations affect understanding, how users respond when AI confidence and accuracy vary, and how interfaces can help people decide when to rely on, question or override AI recommendations.
The work builds on a programme integrating satellite observations, environmental sensors, climate data and AI for forecasting water-quality hazards in Lake Victoria.

The work builds on a wider interdisciplinary programme with the Earth Observation and Environmental Sensing Hub (Dr. Tebbs) developing near-real-time forecasting of water-quality hazards using Earth Observation, low-cost sensors and AI.

Explaining robotic planning decision points along execution

Supervisor: Gerard Canal

Project Description

Explanation of robotic behaviours has been proved to be very important to improve the understanding of the users of such robots, which improves their trust in the robotic system.

However, explanations in robotics are tricky as they need to be given at the correct moment and based on what happened in the execution. In robotic-based planning, an interesting explanation is that of decision points, where the robot could have taken a different action with a different outcome.

This project focuses on the explanation of such decision points at execution time, integrating information om the current events and past events that may help explain the decision to a user. For this, we will look into explainability in the space of plans where, knowing the committed plan and what has happened in the execution, we compare with the other alternatives that the robot had at a certain decision point. This will evolve towards generating explanations along the execution of plans, as well as determining when some decisions may not be obvious to the user, thus warranting explanations.

References

[1] Canal, G., Torras, C., & Alenya, G. (2023). Generating predicate suggestions based on the space of plans: an example of planning with preferences. User modeling and user-adapted interaction, 33(2), 333-357.
[2] Eifler, R., Cashmore, M., Hoffmann, J., Magazzeni, D., & Steinmetz, M. (2020, April). A new approach to plan-space explanation: Analyzing plan-property dependencies in oversubscription planning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 06, pp. 9818-9826).
[3] Wachowiak, L., Celiktutan, O., Coles, A., & Canal, G. (2023, June). A Survey of Evaluation Methods and Metrics for Explanations in Human—Robot Interaction (HRI). In ICRA2023 Workshop on Explainable Robotics.
[4] Wachowiak, L., Tisnikar, P., Canal, G., Coles, A., Leonetti, M., & Celiktutan, O. (2022, August). Analysing eye gaze patterns during confusion and errors in human—agent collaborations. In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 224-229). IEEE.

Verbal reasoning with Bayesian networks

Supervisor: Jeroen Keppens

Project Description

Bayesian networks (BNs) are probabilistic graphical models that allow one to represent uncertain knowledge in domains where variables are interdependent [4]. They are often employed as decision support tools because they combine domain knowledge (often from experts) with probabilistic (statistical) information to compute beliefs (posterior probabilities) about hypotheses or outcomes given observed evidence. This makes them well suited to reasoning under uncertainty: e.g. in medical diagnosis (where symptoms are evidence, diseases are hypotheses), risk assessment, fault diagnosis, and ecological management. They enable not only predictive inference (from causes to effects) but also diagnostic inference (from observed effects back to likely causes), and inter-causal reasoning (how evidence bearing on one cause affects the probabilities of other competing causes).

In decision support, BNs help decision makers by quantifying uncertainty, enabling "what if" analyses (e.g. what if certain evidence is observed or what if certain interventions are made), and computing posterior probabilities that feed into utility or cost-benefit analyses or influence diagrams (an extension of BNs to include decision and utility nodes) to choose among actions.

Application of BNs to decision support problems is not trivial. Construction and validation of BNs typically requires expert knowledge of the modelling technique as well as expert domain knowledge. Moreover, even though BNs compute the rational implications as specified by the models, the results can be counterintuitive to human decision makers due to cognitive biases. A range of explanation techniques have been proposed, but these too require a solid background in BN models [1,3].

This project aims to develop hybrid symbolic and sub-symbolic techniques to produce explainable decision support information using BNs via a natural language interface. To achieve this, existing techniques to generate BN models from first principles [2] will be adapted and extended to enable the construction of BNs from first principles from natural language input via an LLM. BN explanation techniques will be adapted and extended to generate explanation via an LLM.

References

[1] Derks, I. P., & de Waal, A. (2021). A Taxonomy of Explainable Bayesian Networks. arXiv preprint arXiv:2101.1184.
[2] Keppens, J., Shen, Q., & Price, C. (2011). Compositional Bayesian modelling for computation of evidence collection strategies. Applied Intelligence, 35(1), 134-161.
[3] Lacave, C., & Diez, F. J. (2002). A Review of Explanation Methods for Bayesian Networks. The Knowledge Engineering Review, 17(2), 107-127.
[4] Pearl, J. (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann.

Adaptive Human-AI Collaboration for Reasoning and Learning

Supervisor: Tasmina Isalm

Project Description

AI is increasingly used as a collaborator in learning and decision making, providing answers, explanations and recommendations. However, an AI that helps learners reach the correct answer may not necessarily help them develop their own reasoning skills. Providing answers too readily, for example, may reduce opportunities for critical evaluation, discussion and productive disagreement.

This project will investigate how reasoning and learning develop through human-AI collaboration, and how AI agents can adapt when and how they intervene to support these processes. It will examine interventions such as asking questions, challenging arguments, presenting alternative viewpoints or evidence, and choosing when not to intervene. Key questions include which forms of human-AI interaction promote productive reasoning and learning, how intervention strategies affect participation and decision quality, and whether acquired reasoning skills transfer to subsequent tasks without AI support.

The project will combine computational argumentation, multi-agent and consensus modelling, and machine learning to model evolving human-AI interactions and develop adaptive intervention strategies. Controlled collaborative learning studies will evaluate effects on reasoning, learning, participation, decision quality and independent transfer. The research aims to establish how adaptive AI can support collaborative learning while developing, rather than replacing, human reasoning capabilities.

References

1. Kohan Marzagao, D., Murphy, J., Young, A. P., Gauy, M., Luck, M., McBurney, P. & Black, E. (2018). Team Persuasion. Theory and Applications of Formal Argumentation, 159—174.
2. Sengupta, S., Varma, U., & Islam, T. (2025). Empowering Cybersecurity Education: A Review of Adaptive Learning Paradigms and Practical Implications. In Joint Proceedings of IS-EUD 2025: 10th International Symposium on End-User Development, CEUR Workshop Proceedings, Vol. 3978
3. Ba, S., Shi, X., Wu, S. & Lu, G. (2026). Artificial intelligence agents in computer-supported collaborative learning: A systematic literature review. Computers and Education: Artificial Intelligence, 10, 100579.

Modelling Behavioural Change for Adaptive Biometrics

Supervisor: Tasmina Islam

Project Description

Behavioural biometric systems identify individuals from interaction patterns such as keystroke and mouse dynamics. However, these behaviours evolve as people learn, develop proficiency and adapt to different tasks and contexts. Existing systems may treat such variation as behavioural drift, making it difficult to distinguish changes associated with legitimate skill development from contextual variation, identity change or potentially malicious behaviour.

This project will investigate how AI can disentangle relatively stable identity characteristics from behavioural changes associated with learning and developing proficiency. It will examine how behavioural signatures evolve during skill acquisition, which characteristics remain identity-discriminative, and whether learning-related changes persist or transfer across tasks.

The project will develop new computational methods for factorising sources of behavioural variation and determining when biometric adaptation is appropriate, drawing on representation and continual learning, uncertainty modelling and adversarial robustness. Existing behavioural-biometric datasets will be complemented by purpose-built, consented studies in which participants acquire unfamiliar digital skills over repeated sessions, enabling behavioural change to be related directly to developing proficiency. The research aims to advance understanding of how learning shapes digital behaviour while developing biometric systems that can accommodate legitimate human development without compromising identity assurance or security.

References

1. Nambiar, A. P., & Islam, T. (2025). Transformer-based multimodal biometric authentication using signature and EEG signals. In 2025 5th International Conference on Intelligent Cybernetics Technology and Applications (ICICyTA) (pp. 473—478). IEEE.
2. Wang, Y., & Islam, T. (2023). Addressing privacy and security concerns in online game account sharing: Detecting players using mouse dynamics. In ICPRAM 2023 — Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods, Volume 1 (pp. 864—871).
3. Subash, A., Song, I., Lee, I., & Lee, K. (2026). Adaptability of current keystroke and mouse behavioral biometric systems: A survey. Computers & Security, 160.

Reliable and Robust Inference for AI Safety and Security

Supervisor: Nicola Paoletti

Project Description

Modern AI systems are increasingly deployed in safety- and security-critical settings, where predictions and decisions must remain reliable under distribution shifts, adversarial perturbations, and evolving environments. This is especially challenging when undesirable behaviours are rare, difficult to observe, or deliberately concealed, as highlighted by recent work on deception, alignment faking, and prompt-injection attacks [6-9].

Our group develops methods for reliable inference and uncertainty quantification with rigorous probabilistic guarantees, including conformal prediction, conformal risk control, learn-then-test methods, exchangeability martingales, and e-processes [10-13]. Our current work spans safety monitoring of language models, adversarially robust prediction, online prediction under distribution shift, robust counterfactual explanations [4], and predictive monitoring and control of cyber-physical and multi-agent systems [1-4]. We are also involved in the Coefficient Giving-funded project "Verifiably Robust Conformal Probes", combining latent probing with conformal and robustness guarantees (https://www.kcl.ac.uk/news/project-to-verify-ai-models-with-confidence-awarded-major-grant).

The PhD could develop methods at the intersection of reliable inference, uncertainty quantification, and latent representations, including calibrated latent monitors, sequential monitoring, and robust methods for safety, attack detection, diagnosis, control, routing, or verification.

The project would suit a candidate with a strong background in machine learning, statistics, optimisation, or formal methods.

References

1. Pona, Edoardo, et al. "Calibrate-Then-Delegate: Safety Monitoring with Risk and Budget Guarantees via Model Cascades." arXiv preprint arXiv:2604.14251 (2026).
2. Jeary, Linus, et al. "Verifiably robust conformal prediction for probabilistic guarantees under adversarial attacks." Pattern Recognition 170 (2026): 112051.
3. Skalse, Joar, et al. "Online Conformal Prediction Beyond Feedback." arXiv preprint arXiv:2608.07139 (2026).
4. Bilkhoo, Aman, et al. "CONFEX: Uncertainty-Aware Counterfactual Explanations with Conformal Guarantees." arXiv preprint arXiv:2510.19754 (2025).
5. Arditi, Andy, et al. "Refusal in language models is mediated by a single direction." Advances in Neural Information Processing Systems 37 (2024): 136037-136083.
6. Xu, Hainiu, et al. "PreAct-Bench: Benchmarking Predictive Monitoring in LLMs." arXiv preprint arXiv:2606.09890 (2026).
7. R. Greenblatt et al. "Alignment Faking in Large Language Models". arXiv:2412.14093 2.1 (2024).
8. E. Debenedetti et al. "Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents." Advances in neural information processing systems 37 (2024)
9. A. Rose et al. "Detecting Multi-Agent Collusion Through Multi-Agent Interpretability" arXiv:2604.01151 (2026).
10. A. N. Angelopoulos et al. "Learn then test: Calibrating predictive algorithms to achieve risk control." The Annals of Applied Statistics 19.2 (2025): 1641-1662.
11. A. N. Angelopoulos, and S. Bates. "Conformal prediction: A gentle introduction." Foundations and Trends in Machine Learning 16, no. 4 (2023): 494-591.
12. A. Ramdas, P. Grunwald, V. Vovk, and G. Shafer. "Game-Theoretic Statistics and Safe Anytime-Valid Inference." Statistical Science, 2023.
13. Vovk, Vladimir, et al. "Retrain or not retrain: Conformal test martingales for change-point detection." Conformal and Probabilistic Prediction and Applications. PMLR, 2021.
14. Hu, Qitian Jason, et al. "Routerbench: A benchmark for multi-llm routing system." arXiv preprint arXiv:2403.12031 (2024).
15. Sadhuka, Shuvom, et al. "E-valuator: Reliable agent verifiers with sequential hypothesis testing." arXiv preprint arXiv:2512.03109 (2025).

Testing Methodology for Theorem Provers Ecosystems

Supervisor: Dr Mohammad Ahmad Abdulaziz Ali Mansour, Dr Karine Even Mendoza

Project Description

Proof assistants play a central role in applications, ranging from practical software verification, such as verifying a Linux kernel [1], a C compiler [2], or cryptographic software by Apple [3] and AWS [4], to more purely scientific applications in mathematics, such as verifying the correctness of the proof of Fermat's Last Theorem [5]. In software verification, a proof assistant checks a formal argument that a program satisfies its specification and behaves as intended.

Nonetheless, proof assistants are themselves complex software systems. They integrate numerous components, including an IDE, external proof engines (e.g. SMT solvers), internal proof engines (e.g. term-rewriting systems, unifiers or simplifiers), and a kernel that checks proofs. Bugs in these components, or in the interactions between them, can lead to incorrect or misleading results, crashes, hangs, and other failures that affect the reliability and usability of the overall verification system. Although the trusted core of theorem provers such as Isabelle is designed to be highly reliable, formally verifying an entire proof-assistant ecosystem is difficult because many of its components and interactions do not have complete, human-readable formal specifications. AI-driven verification may help automate parts of the verification process, yet it cannot establish correctness without a specification of what the system is intended to do. Automated testing therefore offers a complementary way to assess reliability without requiring a complete formal model of the whole toolchain.

This PhD project will investigate new automated testing methodologies for proof assistants through a focused study of Isabelle. The project will explore the design and effectiveness of: (1) automated test harnesses for individual proof-assistant components; (2) generators and mutators for highly structured test inputs; (3) component-specific test oracles for identifying unexpected behaviour; and (4) cross-version fuzzing and differential testing for discovering regressions, inconsistencies, and behavioural changes.

Through this work, the PhD candidate will be exposed to ideas and techniques from logic, formal methods, programming languages, compilers, and fuzzing, and will develop the ability to connect these areas when addressing research problems in proof-assistant reliability.

References

[1] Klein G, Elphinstone K, Heiser G, Andronick J, Cock D, Derrin P, Elkaduwe D, Engelhardt K, Kolanski R, Norrish M, Sewell T. seL4: Formal verification of an OS kernel. In Proceedings SOSP 2009 (pp. 207-220).
[2] Leroy, X., 2009. Formal verification of a realistic compiler. Commun. ACM 52, 107—115. https://doi.org/10.1145/1538788.1538814
[3] https://support.apple.com/en-gb/guide/security/sec59b0b31ff/web
[4] https://github.com/awslabs/AutoCorrode
[5] https://leanprover-community.github.io/blog/posts/FLT-announcement/
[6] Nipkow, T., Paulson, L.C., Wenzel, M., 2002. Isabelle/HOL - A Proof Assistant for Higher-Order Logic, Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/3-540-45949-9

Privacy in AI Companion / Assistant Technologies

Supervisor: Kovila Coopamootoo

Project Description

AI systems are increasingly entering emotionally sensitive domains, including companionship, relationship advice, education guidance, mental-health support, and sustained personal conversation through agentic companions or assistants. Their integration of large language models (LLMs) facilitate coherent, human-like conversations, and are becoming more and more autonomous. However, they also raise significant privacy concerns, as well as trust, consent and agency challenges.

The PhD projects can include one or a combination, of the following objectives:

- Investigate data governance (and compliance) for AI companions/agents, and envision/design mechanisms to address gaps
- Investigate and design for privacy (/ trust / consent / agency)
o With a user centric perspective: Understand users, explore design possibilities, co-development
o Enforcement through privacy/security mechanisms implementation
- Assess risks factors and implications, e.g. to online safety, privacy, safety harms, in particular across human—AI interaction (HAI) and impacts on human—human relationships (HHI)
- Investigation of and calibrating human-AI decision-making
o Preferences, vulnerability, risk, design

Biologically Inspired Artificial Intelligence for Understanding and Mitigating Age-Related Cognitive Decline

Supervisor: Kathleen Steinhofel / Peter Giese

Project Description

Age-related cognitive impairment, including declines in memory, learning, and decision-making, is a growing global health challenge. While artificial neural networks (ANNs) have achieved remarkable success in modelling cognitive processes, their learning mechanisms differ fundamentally from those observed in the ageing brain. This PhD project aims to develop biologically inspired AI models based on Long-Term Potentiation (LTP), a key mechanism supporting learning and memory in younger brains, and Multi-Innervated Spines (MIS), neural structures that become more prevalent with ageing and may contribute to adaptive compensation for cognitive decline. By incorporating these mechanisms into artificial neural networks, the research will investigate how changes in synaptic plasticity influence learning speed, memory retention, adaptability, and cognitive resilience. The project seeks to provide new insights into the neural basis of age-related cognitive impairment while developing novel AI architectures capable of more robust, flexible, and energy-efficient learning. Ultimately, the work aims to bridge neuroscience and artificial intelligence, contributing both to our understanding of brain ageing and to the design of next-generation brain-inspired intelligent systems.

References

Simulating Categorical Memory Engram Formation Under Young and Ageing Brain Conditions
J Hewavitharana, PHDC Avelar, P Giese, K Steinhofel
Artificial Life Conference Proceedings 38 2026 (1), 12, 2026
Low concentrations of amyloid-beta oligomers induce synaptogenesis characteristic for mild cognitive impairment and alter the de novo proteome
K Wu, S Lee, R Martinez-Serra, L Zhang, S Lynham, KP Giese
Translational psychiatry 16 (1), 132, 2026
Aging-independent decrease of complex multi-spine boutons in hippocampal area CA1 after contextual fear conditioning
R Martinez-Serra, S Lee, I Kraev, KP Giese
Molecular brain 19 (1), 1, 2025
Brain Inspired Learning for Neural Networks
J Hewavitharana, A Anand, P Giese, C Moretti Ierardi, K Steinhofel
International Conference on Engineering Applications of Neural Networks, 59-72, 2025
Computational analysis of learning in young and ageing brains
J Hewavitharana, K Steinhofel, KP Giese, CM Ierardi, A Anand
Frontiers in Computational Neuroscience 19, 1565660, 2025

Data Analysis and Optimization for Bicycle Infrastructure Network Design and Operation

Supervisor: Dimitrios Letsios

Project Description

Bicycles are considered one of the main contemporary alternatives to fossil fuel vehicles for reducing carbon emissions in urban environments. They are an integral part of transport decarbonization strategies (e.g. the UK Net Zero Strategy) and widely adopted as an environmentally friendly transport mode for both passenger (conventional and electric bikes) and urban freight (cargo bikes) transport during the last 5 years.

However, the large-scale adoption of bicycles introduces infrastructure and operational challenges that differ substantially from those associated with conventional motorised transport. From an infrastructure perspective, cycling can be constrained by safety concerns, physical effort, inadequate cycling facilities, and challenging urban geographies. Electric bicycles introduce additional operational requirements, including battery charging and swapping, which may be time-consuming and constrained by the availability and accessibility of charging infrastructure. Furthermore, bicycle networks can exhibit considerable spatial variation in accessibility and connectivity, potentially leading to inefficiencies as well as inequities in access to safe and effective cycling infrastructure.

This project will develop data-driven and optimisation methodologies to support stakeholders, including cyclists, local communities, transport operators, and policymakers, in addressing these challenges and improving the safety, spatial equity, efficiency, and scalability of bicycle networks. To this end, the project will advance methods in spatial data science, operations research, and optimization to assess the bikeability of London's transport network, identify suitable infrastructure interventions, and support the planning and operation of bicycle-based transport systems.

The research will make use of spatial and operational data, together with subject-matter expertise, from a range of sources. These are expected to include geographic databases, a London local authority, bike-sharing operators, and companies providing cargo-bike logistics services. The resulting methodologies will be evaluated using real-world data and computational experiments, with the aim of generating both methodological advances and actionable insights for practitioners and policymakers.

Applicants are expected to have strong quantitative skills in data science, operations research, and machine learning, ideally demonstrated through relevant coursework, research projects, or professional experience. Proficiency in Python will be essential. The PhD will involve substantial work in data processing and analysis, extraction and communication of data-driven insights, mathematical modelling, algorithm design and optimisation, and the evaluation of proposed methods through computational and data-driven approaches.

References

[1] Christoph Steinacker, Mads Paulsen, Malte Schroder, and Jeppe Rich. Robust design of bicycle infrastructure networks. Nature, 2025.
[2] Juan P. Ospina et al. The maximal covering bicycle network design problem. Transportation Research Part A, 2022..

Counterfactual Causal Reasoning for Sequential Decision-Making and AI Systems

Supervisor: Nicola Paoletti

Project Description

Sequential decision-making systems such as MDPs and POMDPs underpin reinforcement learning, robotics and formal verification, but explaining why a particular decision or failure occurred remains challenging. Counterfactual causal reasoning asks what would have happened under alternative interventions, supporting tasks such as counterfactual explanation and actual-cause analysis [1].

Our group develops structural-causal-model methods for these problems. Recent work includes robust counterfactual inference for MDPs under causal-model uncertainty, extensions to data-driven uncertainty, inference for nondeterministic causal models, counterfactuals for language models, and temporal logics with counterfactual modalities [2-5].

Potential directions include:

* Causal attribution and repair for language-model agents, developing efficient algorithms for symbolic MDP abstractions and counterfactual replay. [6-9]
* Latent-level causal explanations, tracking interpretable concepts across multi-step or multi-agent interactions, with applications to explaining and mitigating misaligned AI behaviours. [10-13]
* Counterfactual analysis of learned world models, including Dreamer- and JEPA-style architectures.
* Formal verification of counterfactual properties, including numerical model checking, alternative causal semantics, continuous-state/time systems, and connections to hyperproperties [14-16].

The ideal candidate would have strong foundations in probability, machine learning, formal methods, causal inference, or reinforcement learning, and an interest in rigorous methods for understanding AI systems.

References

1. Halpern, Joseph Y. "A modification of the Halpern-Pearl definition of causality." arXiv preprint arXiv:1505.00162 (2015).
2. Lally, Jessica, et al. "Robust Counterfactual Policy Optimisation via Nondeterministic Causal Models." arXiv preprint arXiv:2608.02893 (2026).
3. Pona, Edoardo, et al. "Abstract Counterfactuals for Language Model Agents." Advances in Neural Information Processing Systems 38 (2026): 87484-87509.
4. Kazemi, Milad, Jessica Lally, and Nicola Paoletti. "Causal temporal reasoning for Markov decision processes." Research Directions: Cyber-Physical Systems 3 (2025): e3.
5. Lally, Jessica, Milad Kazemi, and Nicola Paoletti. "Robust counterfactual inference in Markov decision processes." arXiv preprint arXiv:2502.13731 (2025).
6. Beckers, Sander, Frederick Eberhardt, and Joseph Y. Halpern. "Approximate causal abstractions." Uncertainty in artificial intelligence. PMLR, 2020.
7. Lopez-Miguel, Ignacio D., et al. "ATLAS: Discovering Agent Strategies through LLM-Guided Abstraction and Automata Learning." arXiv preprint arXiv:2608.14352 (2026).
8. Bonagiri, Akash, et al. "CausalFlow: Causal Attribution and Counterfactual Repair for LLM Agent Failures." arXiv preprint arXiv:2605.25338 (2026).
9. Zhang, Shaokun, et al. "Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems." arXiv preprint arXiv:2505.00212 (2025).
10. Mueller, Aaron, et al. "The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis." Computational Linguistics 52.1 (2026): 331-378.
11. Rose, Aaron, et al. "Detecting Multi-Agent Collusion Through Multi-Agent Interpretability." arXiv preprint arXiv:2604.01151 (2026).
12. Geiger, Atticus, et al. "Causal abstractions of neural networks." Advances in Neural Information Processing Systems 34 (2021): 9574-9586.
13. Shi, Chuancheng, et al. "TraceRouter: robust safety for large foundation models via path-level intervention." arXiv preprint arXiv:2601.21900 (2026).
14. Dimitrova, Rayna, Bernd Finkbeiner, and Hazem Torfah. "Probabilistic hyperproperties of Markov decision processes." ATVA, 2020.
15. Von Kugelgen, Julius, Abdirisak Mohamed, and Sander Beckers. "Backtracking counterfactuals." Conference on Causal Learning and Reasoning. PMLR, 2023.
16. Baier, Christel, Jakob Piribauer, and Robin Ziemek. "Foundations of probability-raising causality in Markov decision processes." Logical Methods in Computer Science 20 (2024).

Quantum Machine Learning: Characterising the Quantum Advantage

Supervisor: Mohammad Reza Mousavi (Informatics), Rocio Martinez Nunez (Infectious Diseases)

Project Description

Quantum Machine Learning (QML) has emerged as a novel extension of "traditional" machine learning where parts of the machine learning pipeline are replaced by quantum circuits. Quantum computing processes information using qubits (instead of classical bits, 1 or 0), which can be in any combination of 0 and 1 states simultaneously[RM1.1]. Due to the rich input and output space of quantum computing, this promises novel ways of encoding and processing the data, potentially leading to faster convergence and/or better performance, and different answers to classical methods[RM2.1]. Many of the traditional machine learning techniques have now been extended to use the rich space of quantum information with early results suggesting quantum advantages, e.g., in terms of reduced training time, for the interpretation of complex datasets such as those contained in omics technologies, currently employed to interrogate disease-defining changes at the molecular level.

The aim of this project is to characterise the shape of data that gives various quantum machine learning techniques a practical advantage over classical data. We will start off with simple classification techniques (such as quantum support vector machines) and will move on to more sophisticated architectures (such as quantum neural networks).

We want to characterise the relevance of these approaches in solving biologically meaningful questions. While great advances have been done to discover the genetic make-up of cells, decoding and interpreting what it means in the subsequent steps in gene expression remains a hard task. Machine learning is currently used to classify gene expression patterns, or determine the most important parameters that explain large variation patterns in gene expression. However, the biological relevance of these models remains elusive, and they are likely missing connections that cannot be revealed using classical methods. We aim to understand fundamental questions underpinning gene expression patterns of importance during infection and inflammation. When cells encounter external stimuli, such a virus, they reprogramme themselves to be able to survive infection and stop the virus from spreading and replicating. This entails changes in the levels of thousands of molecules, in different places inside the cells. Current methodologies enable measuring thousands of mediators and their localization, but good and novel tools to easily establish predictive patterns or reduce the dimensionality of datasets to the minimal features that underlie these changes is a hard computational task.

We will investigate the extent of the quantum advantage promise on different machine learning tasks, with different hyper-parameters and measure its effect in a systematic manner. Our project will be conducted on data provided by colleagues from infectious diseases to demonstrate practical advantages in concrete applications in drug discovery and causal discovery in respiratory and inflammatory diseases.
Project aims and objectives

As an application area, we aim is to develop a qML model that enables integration of omics datasets to predict antiviral and inflammatory responses. The student will do this following these objectives:
- Develop a classifier to group molecules into clusters underpinning certain biological outputs (e.g. antiviral status, anti-inflammatory status);
- Develop a qML model that reduces the number of variables to those key underpinning the gene expression output;
- Characterise the boundaries of quantum advantage of Quantum ML techniques over their classical counterparts, particularly focussing on the qML models and the biological datasets developed above.

References

Elias F. Combarro and Samuel Gonzalez-Castillo, A Practical Guide to Quantum Machine Learning and Quantum Optimization: Hands-on Approach to Modern Quantum Algorithms, Packt, 2023
Zoran Krunic, Frederik F. Flother, George Seegan, Nathan D. Eearnest-Noble , and Omar Shehab. Quantum Kernels for Real-World Predictions Based on Electronic Health
Records. IEEE Trans. on Quantum Engineering. 2022.
Md Nadim, Mohammad Hassan, Ashis Kumar Mandal, Chanchal K. Roy. Quantum vs. Classical Machine Learning Algorithms for Software Defect Prediction: Challenges and Opportunities Quantum. Proc. of Q-SE 2025, ACM/IEEE, 2025.
Xinyi Wang, Shaukat Ali, Paulo Arcaini, Narasimha Raghavan Veeraragavan, Jan F Nygard.
Quantum Neural Network Classifier for Cancer Registry System Testing: A Feasibility Study.
arXiv preprint arXiv:2411.04740, 2024.

Black-Box Detection of Neural Network Backdoor Attacks Using Causality

Supervisor: Hana Chockler

Project Description

Neural networks are used in many applications, such as image analysis, autonomous systems, and security. However, they can be vulnerable to backdoor attacks, where an attacker modifies the model internals so that it behaves normally on most inputs but produces incorrect predictions when specific features are present. Since the model appears to function correctly in standard testing, such attacks can be difficult to detect.

Explainable AI (XAI) techniques help users understand why a neural network made a particular prediction. Specifically, black-box explanation methods identify the regions of an input image that are most important for a prediction without requiring access to the model's internal structure. While these methods were designed to improve transparency and interpretability, they may also provide clues about potential attacks.

In this project, we will investigate whether such explanations can also be used to detect backdoored models. Using ReX, a causality-based explanation method, we will generate explanations for image classifications and examine their consistency under different image augmentations. Our hypothesis is that backdoored models rely on decision mechanisms that lead to explanations less consistent than those produced by non-attacked models. The project will evaluate this hypothesis on a range of backdoor attacks and augmentation techniques, with the goal of developing a practical black-box approach for detecting hidden backdoors.

References

Facial Misrecognition Systems: Simple Weight Manipulations Force DNNs to Err Only on Specific Persons
https://arxiv.org/abs/2301.03118

Occlusion-Aware Training for Post-Hoc Neural Network Explainability

Supervisor: Hana Chockler

Project Description

Neural networks are used in many applications, such as image analysis, autonomous systems, and security. However, their decision making process is opaque, giving no insight into how the prediction was made. To address this challenge, researchers have developed post-hoc black-box explainability methods that attempt to explain a model's prediction without requiring access to its internal structure. Many popular techniques create explanations by occluding parts of the input and observing how the model's output changes, thereby estimating which input regions are more important for the prediction. However, occlusion can generate out-of-distribution samples, raising questions about the reliability of the resulting explanations. Furthermore, selecting an appropriate occlusion value often needs domain expertise, and even well-chosen values can still affect the model's predictions.

In this project, we will develop a training objective (loss function) that enforces neural networks to be invariant to a user-defined occlusion value. The training objective can be architecture-agnostic and incorporated into the training process of existing (pretrained) models. By learning to ignore the designated occlusion value during decision-making, the resulting models can be reliably audited using post hoc black-box explanation methods, with a reduced risk of introducing unintended artifacts due to occlusion.

References

Activation-Deactivation: A General Framework for Robust Post-hoc Explainable AI
https://arxiv.org/abs/2510.01038

Scalable Oversight of AI Agents

Supervisor: Yali Du

Project Description

This PhD project will investigate how less capable AI systems can reliably supervise more capable language model agents. As agents undertake increasingly complex tasks, their outputs and explanations may become difficult to verify, creating a need for oversight methods that remain effective across capability gaps.

The research will develop an empirical and methodological framework for understanding when oversight succeeds, why it fails, and how limited supervision resources should be allocated. It will examine three connected questions: how capability differences affect error detection; whether evidence, verification tools, and structured debate improve supervisory judgements; and how selective escalation to stronger overseers can balance reliability and cost. Particular attention will be paid to plausible but misleading explanations and failures that emerge when agents adapt to oversight.

Using tasks with independently verifiable outcomes, the project will combine controlled experiments, modelling of oversight performance, and the development of adaptive supervision methods. Evaluation will measure accuracy, robustness, computational cost, and generalisation across tasks and models.

The expected contribution is a principled understanding of the limits of scalable oversight, alongside practical methods for improving the reliability and efficiency of supervising increasingly capable AI agents.

References

Engels et al. Scaling Laws For Scalable Oversight. https://arxiv.org/abs/2504.18530
Bowman et al. Measuring Progress on Scalable Oversight for Large Language Models. Paper
Bengio. Why are AI agents lying, cheating and coordinating? https://arxiv.org/abs/2211.03540
https://yoshuabengio.org/en/blog/why-are-ai-agents-lying-cheating-and-coordinating

Self-Evolving Multi-Agent LLM Systems: Collaborative Training and Recursive Self-Improvement

Supervisor: Yali Du

Project Description

This PhD project will investigate how systems of interacting large language model (LLM) agents can improve their capabilities through collaborative training and iterative self-improvement. Agents may assume complementary roles as task generators, solvers, critics, and evaluators, producing learning signals that support improvements across successive training cycles.

The research will examine when multi-agent interaction produces reliable capability gains beyond individual self-improvement. Key questions include how agent diversity, communication, and training objectives influence learning; how agents can generate increasingly useful tasks and feedback; and whether improvements to task generation and evaluation enable sustained recursive self-improvement (RSI). The project will distinguish improvements arising from additional inference computation, memory, or workflow adaptation from those achieved through model training.

Methods will combine multi-agent reinforcement learning, synthetic data generation, and adaptive curricula. Controlled experiments will compare collaborative and individual training under matched resource budgets, using held-out tasks and independent evaluation to assess generalisation, efficiency, and stability. Particular attention will be paid to error amplification, reward exploitation, and loss of diversity.

The expected contribution is a framework for understanding and developing self-evolving agent systems, identifying both the mechanisms that enable sustained improvement and the conditions that limit it.

References

1. *MAPoRL: Multi-Agent Post-Co-Training for Collaborative Large Language Models with Reinforcement Learning.* (2025). [arXiv:2502.18439](https://arxiv.org/abs/2502.18439).
2. *MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models.* (2024). [arXiv:2402.01620](https://arxiv.org/abs/2402.01620).
3. Chen, Z., et al. (2024). *Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models.* [arXiv:2401.01335] (https://arxiv.org/abs/2401.01335).
4. Zhao, A., et al. (2025). *Absolute Zero: Reinforced Self-play Reasoning with Zero Data.* [arXiv:2505.03335] (https://arxiv.org/abs/2505.03335).
5. Zhang, J., et al. (2025). *Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents.* [arXiv:2505.22954] (https://arxiv.org/abs/2505.22954).

Scaffolding Student Learning through GenAI in Cybersecurity Education

Supervisor: Hannah Cao / Ievgeniia Kuzminykh

Project Description

Scaffolding, a concept rooted in Vygotsky's Zone of Proximal Development (ZPD), refers to the instructional supports that help learners bridge the gap between their current abilities and learning goals. Traditionally, scaffolding has been provided by tutors, peers, or carefully designed instructional materials.
Generative AI (GenAI) tools such as ChatGPT are now transforming higher education. With the capacity to generate coherent text, explain complex concepts, and solve problems, these tools offer new opportunities for personalised learning. However, their adoption remains controversial. While they can enrich teaching and learning, concerns persist regarding academic integrity and the risk that students may bypass deeper cognitive engagement.

This project responds to these challenges by designing and evaluating a GenAI-driven scaffolding system for cybersecurity education. Rather than supplying direct answers, the system will prompt reflection, break down problems into manageable steps, and encourage exploration, drawing on established pedagogical approaches such as inquiry-based learning. By integrating advances in AI with the principles of effective scaffolding, this project aims to enhance cybersecurity education while contributing to broader debates about the ethical and pedagogical role of AI in higher education.

References

[1] Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press (1978)
[2] Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., Wang, F.L.: Retrieval-augmented generation for educational application: A systematic survey. Computers and Education: Artificial Intelligence 8, 100417 (2025). https://doi.org/10.1016/j.caeai.2025.10041
[3] Liu, M., M'Hiri, F.: Beyond traditional teaching: Large language models as simulated teaching assistants in computer science. In: Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1. pp. 743—749. SIGCSE 2024, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626252.3630789

Validation of Cyber Threat Intelligence Leveraging Large Language Models

Supervisor: Hannah Cao

Project Description

Cyber Threat Intelligence (CTI) is central to modern cybersecurity, equipping organisations to anticipate and mitigate evolving threats. Effective CTI enables proactive defence through timely incident response, patch management, and vulnerability awareness. However, the rapid advancement of Artificial Intelligence (AI), particularly Large Language Models (LLMs), has introduced new challenges. While LLMs offer powerful tools for processing the unstructured data that dominates CTI, they can also be exploited by malicious actors to generate fabricated intelligence at scale. Such disinformation threatens to mislead analysts, poison datasets, and erode the trust essential for collaborative defence.

Existing validation methods, which rely heavily on structured Indicators of Compromise (IoCs) or source reputation, are ill-equipped to handle nuanced, text-based intelligence of uncertain authenticity. This project addresses this critical gap by developing scalable methods to detect fabricated CTI. Specifically, it proposes leveraging LLMs themselves, guided by structured analysis and corroboration against external sources, to assess the likelihood of falsehood in unstructured reports.

The research will deliver a robust framework for validating CTI, strengthening resilience against AI-driven disinformation campaigns. By enhancing confidence in intelligence sharing, the project aims to reduce the risks of misallocated resources and flawed security postures, ultimately supporting more effective cybersecurity practices. Beyond technical validation, the project will also provide an educational platform for cybersecurity practitioners, equipping them with the knowledge and tools to recognise and counteract fabricated CTI in practice.

References

[1] P. Ranade, A. Piplai, S. Mittal, A. Joshi, T. Finin, Generating fake cyber threat intelligence using transformer-based models, 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1—9. doi:10.1109/IJCNN52387.2021.9534192.

[2] S. Mitra, S. Neupane, T. Chakraborty, S. Mittal, A. Piplai, M. Gaur, S. Rahimi, LOCALINTEL: Generating organizational threat intelligence from global and local cyber knowledge, 2025. arXiv:2401.10036.

[3] Janku, S., Xiao, H., & Caris, T. (2025). An Educational and Validation Tool for Cyber Threat Intelligence Leveraging Large Language Models. In 3rd International Workshop on Cyber Security Education for Industry and Academia (CSE4IA 2025)

Towards controllable and interpretable large language model alignment

Supervisor: Hanqi Yan/Yulan He

Project Description

Large language models (LLMs) are increasingly deployed in high-stakes settings, yet current alignment methods, most notably reinforcement learning from human feedback (RLHF), introduce a critical vulnerability: reward hacking. Rather than genuinely learning to be helpful, models may exploit the reward signal to produce outputs that satisfy evaluators without reflecting true knowledge or intent. This manifests in several failure modes including sycophancy, deceptive alignment, and faithfulness failures in chain-of-thought reasoning.

This PhD project pursues two research strands. First, using mechanistic interpretability tools including probing, causal interventions, and gradient-based analysis to identify where and how deceptive tendencies arise inside LLMs. Second, developing practical detection and mitigation methods including lightweight probing-based detectors and inference-time interventions that correct deceptive behaviour without retraining. Methods will be evaluated across high-stakes domains including education, healthcare, and legal assistance, contributing to safer and more trustworthy LLM deployment. This project is ideal for students interested in LLMs, interpretability, and alignment, and offers excellent opportunities for impactful research and publication.

References

[1] When thinking backfires: mechanistic insights into reasoning-induced misalignment
[2] Large Language Models Hack Rewards, and Society
[3] Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference
[4] GRADE: Probing Knowledge Gaps in LLMs through Gradient Subspace Dynamics
[5] PreAct-Bench: Benchmarking Predictive Monitoring in LLMs
[6] Automonitor-bench: Evaluating the reliability of llm-based misbehavior monitor

Verbalisation, Visualisation and Search for Explainable AI Planning

Supervisor: Amanda Coles / Gerard Canal

Project Description

In recent years, there has been a large movement in developing Explainable AI (XAI). Trust and Explainability are key challenges to real-world deployment of AI systems: in order for a system to be used, humans need to understand the reasons why certain decisions were made, which in turn helps them to trust the outputs and use AI systems. In this project, you will work as part of a dynamic and supportive team performing world-renowned research in the exciting area of Explainable AI Planning (XAIP).
AI planning is concerned with thinking before acting: given the initial state of the world, a desired goal and a set of actions that can be performed, an AI Planner is a software system that generates a plan: a sequence of actions that transforms the initial state into one that satisfies the goal. A key strength of planning is domain independence: one planner can be applied in a diverse range of scenarios—e.g. robotics, energy network control and transport logistics—giving huge application potential.
Your research will build on research carried out by our team here at King's in a recently completed $1million research project, "Explaining the Space of Plans". The systems we developed are the first capable of explaining plans and the decisions that lead to them in expressive domains, addressing the major challenge of explaining not only differences between plans but the reasons behind those differences (Krarup et. al. 2024). While highly successful, the output of these systems remains very technical, and your research will focus on extending this technology to improve the explanations generated and on the very important issue of how these explanations can be verbalised and displayed effectively to users, building on the plan verbalisation techniques developed by Dr Canal and his team (Canal et. al. 2022) . Funding has been secured for user studies to be conducted to evaluate the outcome of this project.

Project Aims: Verbalisation, visualisation and improvement of the generation of explanations in explainable AI planning. This project has scope to either be very technically focused (based on complex search algorithms to generate explanations, see e.g. (Krarup et. al. 2024)) or to be visualisation/verbalisation focused, building on the foundations of our user study in (Krarup et. al. 2025), or a mixture of the two depending on the students' interests.

Techniques and Skill Development: Throughout the project, you will develop technical skills in Programming, AI Planning, Explainable AI, Experience in visualisation and HCI, as well as working with users and user studies, and statistical analysis of results. You will also be supported in developing significant skills in academic writing, presentation skills and peer reviewing. Due to the domain-independent nature of planning, you will have the opportunity to work with our industrial partners across a wide range of sectors to apply your research to their respective problem domains.

References

Krarup B., Coles A. J., Gao D., Long D., and Smith D. E. (2025). "How Humans Explain the Difference in the Quality of Plans — A User Study". Proceedings of the Human-Aware and Explainable Planning Workshop at the 2025 International Conference on Automated Planning and Scheduling (ICAPS).
Krarup B., Coles A. J., Long D. & Smith D. E. (2024). "Explaining Plan Quality Differences" (2024), Proceedings of the 2024 International Conference on Automated Planning and Scheduling (ICAPS).
Canal G., Krivic S., Luff P., Coles A. (2022). "PlanVerb: Domain-Independent Verbalization and Summary of Task Plans" Proceedings of the 36th AAAI Conference on Artificial Intelligence.
Krarup, B., Krivic, S., Magazzeni, D., Long, D., Cashmore, M. & Smith, D. E. (2021) "Contrastive explanations of plans through model restrictions" Journal of Artificial Intelligence Research. 72, p. 533-612 80.