# PhD projects (Deep learning)

- Urban AI Shields: GeoAI and Multimodal Data for Predicting Extreme Weather Health Impacts on Vulnerable Urban Populations
- Multi-Agent Learning with Large Language Models
- Art-Deco: Artificial Intelligence for Decentralised Economies
- World Models and Reinforcement Learning for Reliable Robot Autonomy
- Software sustainability analysis and improvement
- Learning for Healthy Human-AI Co-Evolution
- Computational Argumentation for Interactive Explainable AI
- Explainable AI for Biomedical Discovery
- Evolving Cybercrime: Improving Threat Recognition and Response
- AI and finance
- Civic and Responsible AI
- Adaptive Human-AI Collaboration for Reasoning and Learning
- Modelling Behavioural Change for Adaptive Biometrics
- Enabling Accessible Remote Communication Environments
- Embodied Approaches to Assistive Technology
- Reliable and Robust Inference for AI Safety and Security
- 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
- Scalable Oversight of AI Agents
- Self-Evolving Multi-Agent LLM Systems: Collaborative Training and Recursive Self-Improvement
- Towards controllable and interpretable large language model alignment

# 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.

# 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.

# Software sustainability analysis and improvement

Supervisor: Kevin Lano

## Project Description

The project would consider techniques for analysing the software environmental sustainability (in the sense of energy use and energy efficiency) of machine learning systems, or of other kinds of software system with high energy use (such as quantitative financial software) using either rule-based analysis and refactoring, or by the use of deep learning techniques such as LLMs to identify energy use flaws and potential refactorings to improve energy efficiency.

There is the potential for industrial collaboration in this area.

## References

K. Lano et al, "Model-based Software Sustainability Analysis and Improvement", SoSym, 2026.
M. Mehditabar and S. Rajput and T. Sharma, "A Validated Taxonomy on Software Energy Smells", arXiv:2604.04809v1.
H. Siala et al, "Using LLMs for software energy-use analysis and improvement", SusMod '26.

# 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

# 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

# Evolving Cybercrime: Improving Threat Recognition and Response

Supervisor: Tasmina Islam / Kovila Coopamootoo

## Project Description

Cybercrime continually evolves as offenders adapt their language, tactics and modes of interaction to new technologies and user behaviours. The growing use of generative AI may accelerate this process, making emerging threats harder both to detect computationally and to recognise by people.

This project will investigate how changes in malicious communications, such as phishing, scams and social engineering, influence threat recognition and response. Computational methods including natural language processing, machine learning, metadata analysis and temporal modelling will be used to identify how malicious narratives and behavioural patterns evolve over time. Human-centred studies will investigate how people recognise and interpret changing threats, the cues they rely on, common misconceptions, and factors that influence their security decisions.

The project will also investigate how people can be better prepared to recognise unfamiliar and changing threats. It will explore how computational findings about emerging threat patterns can inform cybersecurity learning and develop and evaluate approaches that support transferable threat-recognition and decision-making skills rather than reliance on known examples or fixed warning signs.

## References

1. Islam, T., Becker, I., Posner, R., Ekblom, P., McGuire, M., Borrion, H., & Li, S. (2019). A socio-technical and coevolutionary framework for reducing human-related risks in cyber security and cybercrime ecosystems. In International Conference on Dependability in Sensor, Cloud, and Big Data Systems and
Applications (pp. 277-293).
2. Ng M, Bada M, Coopamootoo KPL ( 2023 ) . What We Do in the Shadows: How does Experiencing Cybercrime Affect Response Actions & Protective Practices? . Conference: 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) vol. 00 , 659 - 672 .
3. J. Li and T. Islam, "Learncyber - Addressing Security and Privacy Concerns: Personalised Learning Tools," 2025 5th International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Yogyakarta, Indonesia, 2025, pp. 467-472,

# 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/.

# Civic and Responsible AI

Supervisor: Martim Brandao

## Project Description

This project is open to a variety of subjects falling under the themes of "Civic AI" and "Responsible Robotics", except those falling under Education, Security or Military applications.

The topic should be one the student is passionate about and has already had some exposure to, but is quite free as long as aligned with the interests of the CRAIL lab (www.civicandresponsibleai.com). Interdisciplinary backgrounds are welcome.

Examples of Civic AI projects:
- New system designs and techniques that apply AI/robotics to solve problems of corporate accountability and human rights organisations
- Leveraging public data and collaborative citizen infrastructure for oversight of AI and robot deployments

Examples of Responsible Robotics projects:
- Algorithms for fair, safe, explainable or accountable robots
- Ergonomics of work with robots and worker-centred robotics and HRI
- Algorithms for sustainable algorithms
- Algorithms for robot accident investigation

# 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.

# Enabling Accessible Remote Communication Environments

Supervisor: Timothy Neate

## Project Description

Remote communication, such as videoconferencing, has become an integral part of daily life. However, it often lacks essential elements like non-verbal cues that are critical for effective communication. This can pose significant challenges for people with language impairments. Building on our previous research into the accessibility of videoconferencing for individuals with aphasia, this PhD will explore new ways to enhance remote communication technologies. By collaborating directly with these communities, you will design and develop tools that make videoconferencing and other online environments more inclusive and supportive for all users.

## References

Timothy Neate, Vasiliki Kladouchou, Stephanie Wilson, Shehzmani Shams, ``Just Not Together'': The Experience of Videoconferencing for People with Aphasia during the Covid-19 Pandemic, Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2022.
Humphrey Curtis, Zihao You, William Deary, Miruna-Ioana Tudoreanu, Timothy Neate, Envisioning the (in) visibility of discreet and wearable AAC devices, Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2023.

# Embodied Approaches to Assistive Technology

Supervisor: Timothy Neate

## Project Description

Non-verbal expression plays a crucial role in everyday communication, whether nodding to indicate agreement or using vocal tone to imply a question. For individuals with language impairments, non-verbal cues are essential for both comprehension and expression. However, most assistive technologies overlook these vital communication methods (see our systematic review). This PhD project will extend our research on wearable devices such as smartwatches, smartbadges, and augmented reality (AR) tools, focusing on innovative solutions for non-verbal communication that work insitu, with communication supported without pre-population of devices. We expect this might be enabled by multimodal large language models and similar technologies, but are relatively technology agnostic. You will collaborate directly with communities who experience language impairments to design technologies that support effective communication in real-world settings.

## References

Humphrey Curtis, Timothy Neate, Carlota Vazquez Gonzalez, State of the art in AAC: A systematic review and taxonomy, Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), 2022.
Humphrey Curtis, Zihao You, William Deary, Miruna-Ioana Tudoreanu, Timothy Neate, Envisioning the (in) visibility of discreet and wearable AAC devices, Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2023.
Humphrey Curtis, Ying Hei Lau, Timothy Neate, Breaking Badge: Augmenting Communication with Wearable AAC Smartbadges and Displays, Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2024.
Humphrey Curtis, Timothy Neate Making smartglasses accessible: perspectives and prototypes from co-design with people with aphasia, Nature Scientific Reports, 2025.

# 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).

# 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.

# 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).

# 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