Artificial Intelligence Projects 2026-27

Contents

Game Theory and Learning in Multi-Agent Systems

Supervisor: Stefanos Leonardos

Areas: Artificial Intelligence, Machine learning / Deep learning, Game theory, Multi-agent systems

(Back to Scholarship Not Allocated)

Project Description

This project is designed for students interested in research at the intersection of game theory, learning dynamics, and multi-agent systems, with applications in economics, machine learning, and artificial intelligence. The aim is to study how complex patterns and behaviors emerge when many agents interact and adapt over time. We will explore phase transitions in strategic interactions, investigate the role of chaos and dynamical systems, and develop or analyze novel learning algorithms. The project will combine tools from mathematics, game theory, and AI to understand coordination, competition, and long-term dynamics in real-world systems. Students will have the opportunity to contribute to both theoretical advances and practical applications, helping to shape the future of intelligent learning systems.

References

1. I. Sakos, S. Leonardos, S. A. Stavroulakis, W. Overman, I. Panageas, G. Piliouras. Beating Price of Anarchy and Gradient Descent without Regret in Potential Games, 12th International Conference on Learning Representations (2024).
2. S. Roesch, S. Leonardos & Y. Du. Selfishness Level Induces Cooperation in Sequential Social Dilemmas, 23rd Conference on Autonomous Agents and Multiagent Systems (2024).
3. Leonardos, S., Sakos, J., Courcoubetis, C. and Piliouras, G. (2023). Catastrophe by Design in Population Games: A Mechanism to Destabilize Inefficient Locked-in Technologies. ACM Trans. Econ. Comput. 11, 1—2, Article 1 (June 2023), 36 pages. doi:10.1145/3583782
4. Leonardos, S., and Piliouras, G. (2022). Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory, Artificial Intelligence, Volume 304, 103653, doi:10.1016/j.artint.2021.103653.
5. Leonardos, S., Piliouras, G., and Spendlove, K. (2021). Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality, in Advances in Neural Information Processing Systems, volume 34, pp. 26318--26331, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2021/file/dd1970fb03877a235d530476eb727dabPaper.pdf.
6. Leonardos, S., Overman, W., Panageas I., and Piliouras, G. (2022). Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games, in International Conference on Learning Representations (ICLR 2022), https://openreview.net/forum?id=gfwON7rAm4.

Exploring Interactive Multi-Dimensional Approaches to Patient Communication in Oral Health Education

Supervisor: Informatics: Dr Alfie Abdul-Rahman & Dr Lin Gui FoDOCS: Dr Melanie Nasseripour & Dr Ana Angelova

Areas: Artificial Intelligence, Machine learning / Deep learning, Human-centred computing (human-computer interaction), Natural Language Processing, Systems (software engineering, programming)

(Back to Scholarship Not Allocated)

Project Description

This joint project between the Department of Informatics and the Faculty of Dentistry, Oral & Craniofacial Sciences (FoDOCS) seeks to transform how patient communication is taught in oral health education. Effective patient-clinician communication is central to encouraging oral hygiene practices, supporting behaviour change, and demonstrating professionalism. However, conventional training methods are resource-intensive, requiring significant time, staff, and cost to deliver authentic clinical practice scenarios.

To address these challenges, the project proposes the design and evaluation of interactive, multi-dimensional learning environments, with a particular emphasis on immersive Virtual Reality (VR). VR offers an unparalleled ability to create engaging, realistic, and repeatable patient encounters where students can practice communication skills in a safe but lifelike clinical setting. Complementary approaches—including text-to-text and voice-to-voice communication systems—will expand accessibility, providing adaptable tools for different learners and educational contexts.

Generative Language Models (GLMs) will be harnessed to generate customised patient cases and conversational responses, ensuring that scenarios adapt dynamically to student inputs. These responsive simulations allow students to repeatedly practice communication tasks such as explaining oral hygiene, negotiating dietary changes, or discussing treatment options, thereby reinforcing learning and professional skill development.

Research Questions
- How can immersive VR be designed to simulate realistic patient-clinician interactions that enhance students' communication, empathy, and professionalism?
- What are the comparative benefits and limitations of VR, text-based, and voice-based simulation approaches in oral health education?
- In what ways can GLMs be integrated to generate personalised, contextually relevant, and adaptive patient scenarios that respond in real time to learners' actions?
- How do students perceive and engage with VR-based versus non-immersive approaches, and how does this affect learning outcomes?
- What metrics and evaluation frameworks can be developed to measure improvements in communication competence, confidence, and behaviour change skills using immersive and multi-dimensional tools?
- How can these approaches be scaled sustainably across curricula and adapted to different cultural and institutional contexts in oral health education?

Expected Contributions
- Development of a VR-based immersive training system for patient communication in oral health education.
- Comparative evaluation of VR, text, and voice-based simulations, identifying best practices for different learning settings.
- Integration of GLMs for adaptive, customised case study generation.
- A multi-dimensional toolkit for cost-effective, scalable communication training.
- Evidence-based recommendations for embedding immersive and interactive communication training within oral health curricula.

Energy-Efficient Deep Learning with Sparse Subnetworks

Supervisor: Frederik Mallmann-Trenn

Areas: Artificial Intelligence, Machine learning / Deep learning, Foundations of computing (algorithms, computational complexity)

(Back to Scholarship Not Allocated)

Project Description

Supervisor: Dr Frederik Mallmann-Trenn
Host: Algorithm & Data Analysis (ADA) Group, King's College London
Theme: Energy-efficient machine learning, neural network sparsification, randomized processes
Keywords: supermasks / Edge-Popup, lottery-ticket hypothesis, sparsity, PyTorch/TensorFlow, energy measurement

Background:
Modern AI models are powerful—but they're also expensive in energy. That means higher cloud costs, bigger carbon footprints, and shorter battery life on phones and wearables. One promising route to cut energy is to run much smaller "sparse" networks that keep accuracy high while doing far less computation. This project studies a striking idea: you can often find a high-performing subnetwork inside a randomly initialised model by learning only a binary mask ("supermask")—no heavy weight training required. If we can understand and systematise this, we can make AI cheaper, greener, and more accessible.

Starting reference:
Ramanujan et al., What's Hidden in a Randomly Weighted Neural Network? (CVPR 2020)
https://openaccess.thecvf.com/content_CVPR_2020/html/Ramanujan_Whats_Hidden_in_a_Randomly_Weighted_Neural_Network_CVPR_2020_paper.html

Project overview

You will combine theory (probability and randomized processes) with hands-on engineering to reproduce, extend, and understand supermask methods (also called Edge-Popup).

Objectives:

1) Reproduce and clarify the baseline
2) Implement supermask/Edge-Popup for standard architectures (MLPs, CNNs).
3) Match key results from the paper; run careful ablations (how we parameterise masks, sparsity levels, initialisation, training schedules).
4) Make it practical and energy-aware
5) Explore structured sparsity (e.g., channel/block patterns) that hardware can exploit.
6) Explain when it works (theory)
7) Develop conditions for the existence of good sparse subnetworks under random initialisation (as functions of width, depth, and target sparsity k).
8) Study recoverability: when can simple procedures reliably find such subnetworks?
9) Map trade-offs (accuracy vs sparsity; compute vs expressivity), using tools from concentration of measure, random matrices/graphs, and randomized algorithms.

What you'll gain
1) Expertise in energy-efficient deep learning with reproducible engineering.
2) Deep understanding of randomized processes in modern ML.

Required background (must-have)

1) Mathematics: very solid probability/randomized processes (concentration inequalities, asymptotics), linear algebra, and optimisation.
2) Programming: strong PyTorch or TensorFlow skills (training loops, data pipelines, experiment tracking).
3) Ability to write clean, tested, reproducible research code (git; clear experiment logs).

References

Ramanujan et al., What's Hidden in a Randomly Weighted Neural Network? (CVPR 2020)
https://openaccess.thecvf.com/content_CVPR_2020/html/Ramanujan_Whats_Hidden_in_a_Randomly_Weighted_Neural_Network_CVPR_2020_paper.html

Implementing Differential Privacy in Neural Networks to Enhance Data Security and Anonymization

Supervisor: Frederik Mallmann-Trenn

Areas: Artificial Intelligence, Machine learning / Deep learning, Foundations of computing (algorithms, computational complexity)

(Back to Scholarship Not Allocated)

Project Description

Abstract: This PhD project aims to address the increasing need for robust privacy-preserving mechanisms in machine learning, particularly focusing on the application of differential privacy within neural networks. With the pervasive use of deep learning in processing sensitive information, there is a critical need to develop techniques that can protect individual data points from being reverse-engineered or identified. This research will explore innovative methods to integrate differential privacy into neural network architectures, ensuring the confidentiality of training datasets while maintaining the utility of the models.

Introduction: As neural networks become more ingrained in handling sensitive data, the potential for privacy breaches escalates. Differential privacy provides a framework to quantify and control the privacy loss incurred when releasing information about a dataset. This project will delve into the optimization of differential privacy in neural networks, balancing the trade-off between privacy protection and the predictive performance of the models.

Objectives: To conduct a comprehensive literature review on current approaches and challenges of applying differential privacy in neural networks.
To develop a theoretical framework for differential privacy that is specifically tailored to neural network applications.
To design, implement, and evaluate new algorithms that integrate differential privacy into neural network training processes without significantly degrading model accuracy.
To create a benchmark dataset and evaluation metrics for assessing the performance of privacy-preserving neural networks.
To investigate the impact of differential privacy on various neural network architectures and learning tasks, such as classification, regression, and generative models.
Methodology:
The project will utilize a combination of theoretical, experimental, and empirical methods. Initial efforts will focus on the theoretical underpinnings of differential privacy and its mathematical integration into neural network algorithms. Following this, experimental simulations using synthetic and real-world datasets will be conducted to assess the viability and performance of the proposed models. Empirical validation will be performed by comparing the new models with state-of-the-art privacy-preserving techniques.

It is absolutely necessary to have a strong math and stats background.

References

Dwork, C. and Roth, A., 2014. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3—4), pp.211-407.
https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf
Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L., 2016, October. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 308-318).
https://arxiv.org/pdf/1607.00133.pdf

Explaining robotic planning decision points along execution

Supervisor: Gerard Canal

Areas: Artificial Intelligence, Robotics

(Back to Scholarship Not Allocated)

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.

Goal-based explanations for autonomous systems and robots

Supervisor: Gerard Canal

Areas: Artificial Intelligence, Robotics

(Back to Scholarship Not Allocated)

Project Description

Autonomous systems such as robots may become another appliance found in our homes and workplaces. In order to have such systems helping humans to perform their tasks, they must be as autonomous as possible, to prevent becoming a nuisance instead of an aid.

Autonomy will require the systems or robots to set up their own agenda (in line with the tasks they are meant to do), defining the next goals to achieve and discarding those who can't be completed. However, this may create misunderstandings with the users around the system, who may expect something different from the robot.

Therefore, it is important that these autonomous systems are able to explain why they achieved one task and not another, or why some new (unexpected) task was achieved that was not scheduled. Other sources of misunderstandings may come from action failures and replanning, where the robot finds a new plan to complete an on-going task. In this case, the new plan may be different to the original one, thus changing the behavior that the robot was performing.

This project will explore how to generate goal-based explanations for robots in assistive/home-based scenarios, extracted from goal-reasoning techniques. It will also look at plan repair to enforce cohesion after a replanning to ideally increase the trust and understanding of the users about the system. Those explanations should also contemplate unforeseen circumstances, therefore explaining things based on "excuses" that the robot may give to the user. Finally, we will investigate how to obtain and provide those explanations at execution time, so explaining on-the-go. The methods developed shall be integrated into a robotic system, in an assistive/service robot scenario.

In addition to the available support by the CDT, the candidates will have the opportunity of contributing to the REXAR (UK) and COHERENT (international) research projects, while collaborating with and being supported by a network of researchers in aligned areas. These projects focus on reasoning for autonomous robots in assistive scenarios, dealing with explanations at different levels of the robotics system, and reasoning about goals and plans.

References

[1] Canal, G., Borgo, R., Coles, A., Drake, A., Huynh, T. D., Keller, P., Krivic, S., Luff, P., Mahesar, Q-A., Moreau, L., Parsons, S., Patel, M., & Sklar, E. (2020). Building Trust in Human-Machine Partnerships. Computer Law & Security Review, 39.
[2] Hawes, N., Burbridge, C., Jovan, F., Kunze, L., Lacerda, B., Mudrova, L., ... & Hanheide, M. (2017). The strands project: Long-term autonomy in everyday environments. IEEE Robotics & Automation Magazine, 24(3), 146-156.
[3] Aha, D. W. (2018). Goal reasoning: Foundations, emerging applications, and prospects. AI Magazine, 39(2), 3-24.
[4] Bercher, Pascal, et al. "Plan, repair, execute, explain—how planning helps to assemble your home theater." Proceedings of the International Conference on Automated Planning and Scheduling. Vol. 24. No. 1. 2014.
[5] Chakraborti, Tathagata, Sarath Sreedharan, and Subbarao Kambhampati. "The emerging landscape of explainable AI planning and decision making. IJCAI 2020.
[6] Gobelbecker, M., Keller, T., Eyerich, P., Brenner, M., & Nebel, B. (2010, April). Coming up with good excuses: What to do when no plan can be found. In Proceedings of the International Conference on Automated Planning and Scheduling (Vol. 20, No. 1).

Adaptation and effective communication in collaborative physically Assistive Tasks

Supervisor: Gerard Canal

Areas: Artificial Intelligence, Robotics

(Back to Scholarship Not Allocated)

Project Description

Physical robotic Assistance can often be modelled as a collaborative task in which the goal of both the user and the robot is to complete an assistive task together. However, assistive settings have a lot of particularities that differentiate them from traditional Human-Robot Collaboration tasks.

For it to be effective, the assistance should be seamless, natural, and without a required effort on the user's side. This means that these robots must be able to communicate with the user in a very natural and intuitive way, but also in an adaptive manner.

In this project, we will investigate the development of techniques for the online adaptation of the robot to the human, as well as anticipation of user needs, and seamless communication in the context of assistive tasks such as robotic feeding and dressing.

References

[1] Canal, G., Alenya, G., & Torras, C. (2019). Adapting robot task planning to user preferences: an assistive shoe dressing example. Autonomous Robots, 43(6), 1343-1356.
[2] Canal, G., Alenya, G., & Torras, C. (2016). Personalization framework for adaptive robotic feeding assistance. In Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8 (pp. 22-31). Springer International Publishing.
[3] Bhattacharjee, T., Lee, G., Song, H., & Srinivasa, S. S. (2019). Towards robotic feeding: Role of haptics in fork-based food manipulation. IEEE Robotics and Automation Letters, 4(2), 1485-1492.
[4] Bhattacharjee, T., Gordon, E. K., Scalise, R., Cabrera, M. E., Caspi, A., Cakmak, M., & Srinivasa, S. S. (2020, March). Is more autonomy always better? exploring preferences of users with mobility impairments in robot-assisted feeding. In Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction (pp. 181-190).
[5] Ondras, J., Anwar, A., Wu, T., Bu, F., Jung, M., Ortiz, J. J., & Bhattacharjee, T. (2022, August). Human-robot commensality: Bite timing prediction for robot-assisted feeding in groups. In 6th Annual Conference on Robot Learning.

Assessing the value of evidence with argument-driven credal networks

Supervisor: Jeroen Keppens

Areas: Artificial Intelligence

(Back to Scholarship Not Allocated)

Project Description

Criminal investigations and court proceedings are (or should be) evidence-led activities. They seek to assess the value of available and yet to be collected evidence in its ability to discriminate between lines of inquiry, or between defence and prosecution hypotheses [3].

There are at least three distinct groups of techniques to assess the value of evidence. Narrative approaches aim to provide explanations that supports the evaluation of coherence of evidence [5]. Argument-driven approaches scrutinise and explain inferences associated with evidence [4]. Bayesian approaches infer new information by calculating the rational, probabilistic implications of one's beliefs [1]. Thus, different types of approach evaluate distinct aspects of evidence and reasoning about evidence. Hybrid approaches have been proposed to combine such distinct aspects.

One type of hybrid approach proposes the specification of argumentation models to define constraints on the node probability tables in a Bayesian network model of evidential reasoning [3]. Bayesian models rely on node probability tables to represent first principles, knowledge inferred from data, and expert opinions. However, if the underlying models are incorrect, the information inferred by such module is inherently unreliable. Node probability tables are inherently difficult to be validated, especially by legal professionals, domain experts, judges, and juries who usually lack familiarity with Bayesian network models. However, if the node probability tables are derived from argumentation models, the argumentation models provide an interface for this scrutiny.

Constraints on node probability tables do not normally lend themselves to the definition of node probability tables as they are used in Bayesian networks. Credal networks are a generalisation of Bayesian networks, where the conditional probabilities in node probability tables are defined by sets [2]. This project aims to extend previous work to define argumentation models of constraints on node probability tables, to define credal networks from said constraints. It also aims to extend Bayesian techniques to assess the value of evidence to credal network, and explain how outcomes depend on arguments.

References

[1] R. Cook, I. Evett, G. Jackson, P. Jones, and J. Lambert. A model for case assessment and interpretation. Science and Justice, 38(6):151—156, 1998.
[2] F. Cozman. Credal networks. Artificial Intelligence 120(2): 199—233, 2000.
[3] J. Keppens. On modelling non-probabilistic uncertainty in the likelihood ratio approach to evidential reasoning. Artificial Intelligence and Law, 22(3):239—290, 2014.
[4] H. Prakken. Analysing reasoning about evidence with formal models of argumentation. Law, Probability and Risk, 3(1):33—50, 2004.
[5] C. Vlek, H. Prakken, S. Renooij, and B. Verheij. Representing the quality of crime scenarios in a bayesian network. In Proceedings of the 28th International Conference on Legal Knowledge and Information Systems, pages 131—140, 2015.

Computational Argumentation for Interactive Explainable AI

Supervisor: Antonio Rago

Areas: Artificial Intelligence, Machine learning / Deep learning, Human-centred computing (human-computer interaction), Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Context:
Computational argumentation has been shown to be an effective formalism for extracting the relevant information from AI models and providing faithful, transparent and interactive explanations to users in a variety of formats, such as those which are conversational [1]. Argumentative explanations have been shown to be useful when applied different AI models, including large language models [2], classification methods [3,4] and recommender systems [1,5]. Recently, a general framework based on argumentation has been introduced [6] that is claimed to be able to not only model interactive explanations provided by AI models, but also the reasoning processes and cognitive biases in humans. However, these claims have currently been verified only by theoretical analysis and evaluations based on simulations, without the user studies which would provide conclusive evidence of the approach's capabilities.

Aims and Methodology:
The aim of this project is to investigate the ability of argumentative approaches, perhaps inspired by [6], to model human reasoning in interactive explanations for a range of AI models in different applications. The student will define instances of argumentative technologies as methods for explaining AI models, e.g. large language models or image classifiers. The methods will then be implemented, giving novel tools for interactive explanations which shifts the decision making towards the user, as in novel Evaluative AI paradigm [7]. The method will then be evaluated by means of machine-centric properties, such as faithfulness, using publicly available datasets in domains such as engineering and healthcare. The project will then explore the evaluation by user-centric properties of explainable AI, such as user satisfaction or trust, as in [1]. The student will then undertake user studies, which may be crowdsourced via Prolific (https://www.prolific.com). The study of interactive explanation for AI models is fertile ground, not least with the rise of large language models, and we believe that this project could have significant impact on the research landscape, particularly given the cutting-edge proposals of [7].

Requirements:
An interest (or, preferably, experience in) in the theory of computational logic and argumentation will be essential, as well as the implementation and analysis of machine learning models and explainable AI methods. This work will follow much of the argumentation-related work shown here: https://antoniorago.github.io/publications/

References

References:
[1] Argumentative Explanations for Interactive Recommendations; Rago et al.; AIJ 2021.
https://www.sciencedirect.com/science/article/pii/S0004370221000576
[2] Argumentative Large Language Models for Explainable and Contestable Claim Verification; Freedman et al.; AAAI 2025.
https://doi.org/10.1609/aaai.v39i14.33637
[3] Data-Empowered Argumentation for Dialectically Explainable Predictions; Cocarascu et al.; ECAI 2020.
http://ebooks.iospress.nl/publication/55172
[4] A Little of That Human Touch: Achieving Human-Centric Explainable AI via Argumentation; Rago; IJCAI 2024.
https://www.ijcai.org/proceedings/2024/983
[5] Argumentation-Based Recommendations: Fantastic Explanations and How to Find Them; Rago et al.; IJCAI 2018.
https://www.ijcai.org/proceedings/2018/269
[6] Interactive Explanations by Conflict Resolution via Argumentative Exchanges; Rago et al.; KR 2023. https://proceedings.kr.org/2023/57/
[7] Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI; Miller; FAccT 2023.
https://dl.acm.org/doi/10.1145/3593013.3594001

Verbal reasoning with Bayesian networks

Supervisor: Jeroen Keppens

Areas: Artificial Intelligence, Natural Language Processing

(Back to Scholarship Not Allocated)

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.

Mechanism Design for Robust AI Alignment

Supervisor: Carmine Ventre

Areas: Foundations of computing (algorithms, computational complexity), Artificial Intelligence

(Back to Scholarship Not Allocated)

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 three problematic assumptions. It assumes AI agents are perfectly rational. It can only implement a limited class of functions due to known impossibility results.

The idea is to study novel solution concepts, inspired by the literature on behavioural economics. This will build on the landscape developed by the PI 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. This directly addresses the limitations above by guaranteeing alignment for agents that may refrain from some non-obvious manipulations.

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

Areas: Artificial Intelligence

(Back to Scholarship Not Allocated)

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/.

Conflict Detection and Reconciliation Across Diverse Clinical Knowledge Sources in Pregnancy Care

Supervisor: Yulan He / Runcong Zhao

Areas: Artificial Intelligence, Machine learning / Deep learning, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Pregnancy care draws on diverse guidelines, hospital protocols, and patient information, but this wealth of evidence also creates challenges for women seeking clear, reliable answers. Guidance can be fragmented, inconsistent, or even contradictory, leaving patients uncertain and clinicians struggling to provide consistent advice. At the same time, large language models (LLMs) and conversational AI tools are becoming increasingly common in digital health, yet current systems largely provide generic responses and fail to address the complexities of conflicting information.

Existing automated approaches to guideline conflict detection are limited, often relying on model-level "black box" decisions. Such approaches suffer from several issues: they may favour whichever recommendation appears most frequently, or give undue weight to options mentioned at the beginning or end of a document. Beyond accuracy, this kind of opaque decision-making undermines patient trust, as it provides no clear rationale for why one recommendation is preferred over another.

Our project is therefore novel in advancing fundamental methods for: formally representing heterogeneous guideline knowledge; algorithmically detecting inconsistencies; evaluating and weighting conflicting evidence; and designing explainable reconciliation strategies suitable for patient use. By shifting the focus from simple question answering towards conflict-aware, explainable reasoning, this project addresses a key technical and clinical gap.

The overarching aim is to develop a transparent and trustworthy QA framework that can provide reliable answers even when knowledge sources disagree. Specifically, the project will (1) create formal representation models for pregnancy-related guidance, (2) design algorithms to detect and categorise conflicts, (3) develop evidence weighting and reconciliation strategies, and (4) integrate explainable mechanisms into a patient-facing QA system. Pregnancy care is chosen as the initial testbed given its direct impact on maternal and fetal outcomes, but the proposed methods are domain-agnostic and can be applied across healthcare more broadly.

References

Braun, T., Rothermel, M., Rohrbach, M., and Rohrbach, A. (2025). Defame: Dynamic evidence-based fact-checking with multimodal experts. In: International Conference on Machine Learning (ICML).
Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. (2024). Medagents: Large language models as collaborators for zero-shot medical reasoning. In Findings of the Association for Computational Linguistics (ACL).
Venktesh, V. and Setty, V. (2025). Factir: A real-world zero-shot open-domain retrieval benchmark for fact-checking. In: Proceedings of the ACM Web Conference 2025 (WWW). Resource Track.
Wu, H., Zeng, Q., and Ding, K. (2024). Fact or facsimile? evaluating the factual robustness of modern retrievers. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM).

Bayesian Meta-Reasoning in Large Language Models

Supervisor: Yulan He / Hanqi Yan

Areas: Artificial Intelligence, Machine learning / Deep learning, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

State-of-the-art large language models (LLMs) are typically pre-trained through statistical next-token prediction and often post-trained with reinforcement learning (RL) using reward models. While these methods enable strong performance on reasoning tasks such as mathematics and coding, they face key limitations such as: (1) sensitivity to task frequency and input perturbations; (2) dependence on task-specific annotations and preference data, which limits scalability; (3) unreliable self-generated feedback that fails to capture shared reasoning patterns. These constraints arise because LLMs are optimised to solve tasks directly rather than to learn how to reason. For better generalisation, we argue that models need to instead actively engage in learning-to-reason or meta-reasoning processes, and develop the ability to monitor, regulate, and adapt their reasoning strategies when needed.

This project aims to implement the Bayesian meta-reasoning framework we described in our position paper in ICML 2025. The new framework will enable models to move beyond task-specific optimisation towards adaptive reasoning strategies that generalise across domains. The research will first develop unified benchmarks and metrics to evaluate meta-reasoning across domains, moving beyond accuracy to measures of calibration error, logical consistency, and cross-domain generalisation. It will then design adaptive architectures that dynamically combine reasoning skills, using approaches such as Mixture-of-Experts and Bayesian inverse planning. Self-play will be explored to discover scalable and multifaceted rewards without costly human annotations. Finally, latent-space reasoning methods will be explored to reduce cascading errors and improve efficiency. Techniques such as diffusion-based reasoning and looped transformers will be explored. If time permits, the project may extend to interpretable meta-knowledge consolidation, using mechanistic interpretability to link reasoning skills to model components. This would allow selective fine-tuning and more efficient training.

References

Yan, H., Zhang, L., Li, J., Shen, Z. and He, Y., 2025. Position: LLMs Need a Bayesian Meta-Reasoning Framework for More Robust and Generalizable Reasoning. In 2025 International Conference on Machine Learning (ICML).

Towards controllable and interpretable large language model alignment

Supervisor: Hanqi Yan/Yulan He

Areas: Artificial Intelligence, Machine learning / Deep learning, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Large language models (LLMs) are rapidly transforming how we interact with technology, but making them aligned with human values and needs remains a major open challenge. Current post-training methods (e.g., reinforcement learning from human feedback) can make models more helpful, accurate, or supportive. However, they face important limitations: (i) Models often need to balance multiple objectives at once (e.g., being safe, empathetic, and logically sound). (ii) Alignment needs can change depending on the person and context. (iii) Models risk forgetting prior abilities when retrained for new goals.

This PhD project will explore how to make LLM alignment both controllable (so we can flexibly adjust a model's behavior) and interpretable (so we can understand why the model behaves as it does). The research will have two main strands: (i) Understanding Alignment Mechanisms: use mechanistic interpretability tools (e.g., sparse autoencoders, probing, causal interventions) to study how alignment signals are represented or entangled inside LLMs (ii) Building Controllable Alignment Frameworks. It is to develop adaptive methods that integrate multiple alignment goals. Explore inference-time control, such as representation steering to adjust model behavior dynamically in response to user queries. The framework will be applied to real-world challenges such as mental health chatbots, where LLMs must combine safety, reasoning, empathy, and clear communication. The broader impact will be methods for more transparent, reliable, and adaptable LLMs. This project is ideal for students interested in LLMs, interpretability, and alignment, and offers excellent opportunities for impactful research and publication.

References

[1] Thinking Hard, Going Misaligned: Emergent Misalignment in LLMs
[2] Sparse Activation Editing for Reliable Instruction Following in Narratives
[3] Soft Reasoning: Navigating Solution Spaces in Large Language Models through Controlled Embedding Exploration
[4] Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference
[5] NOVER: Incentive Training for Language Models via Verifier-Free Reinforcement Learning

Adaptive and Inclusive Cybersecurity Education

Supervisor: Tasmina Islam

Areas: Artificial Intelligence, Machine learning / Deep learning, Cybersecurity, Human-centred computing (human-computer interaction), Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Cybersecurity awareness is essential in a digital society, yet current initiatives are often generic and fail to address the needs of underrepresented groups such as children, minority communities, and people with disabilities. This project proposes to design adaptive awareness environments that apply artificial intelligence to deliver personalised, culturally relevant, and accessible cybersecurity training.

Participants will engage with interactive simulations of cyber threats (e.g., phishing, ransomware, data leakage) enhanced by generative AI, which can create customised attack narratives and provide context-aware, real-time feedback. Alongside the development of adaptive systems, the project will employ interviews and focus groups, combined with behavioural analytics and surveys, to understand user needs, evaluate effectiveness, and identify different types of cybersecurity users (e.g., risk-takers, cautious users, or overconfident participants). It will also explore new ways of measuring awareness outcomes, introducing resilience-based metrics such as detection speed, recovery performance, and transfer of skills across scenarios.

By combining technical innovation with human-centred research, this work seeks to establish a scalable framework for evidence-driven cybersecurity awareness, empowering diverse populations to navigate digital risks safely. Prospective students can discuss options with the supervisor.

References

1. Segupta, 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, 16-18 June 2025, Munich, Germany.: 3rd International Workshop on Cyber Security Education for Industry and Academia.
2. Hedges, M., & Islam, T. (2024). VirSec - Immersive Security Training within Virtual Reality. In 17th International Conference on Advanced Visual Interfaces: 2nd International Workshop on CyberSecurity Education for Industry and Academia (CSE4IA 2024)
3. Islam, T & Zou, Y 2023, ChildSecurity: A Web-based Game to Raise Awareness of Cybersecurity and Privacy in Children. in Cybersecurity Challenges in the Age of AI,Space Communications and Cyborgs.

Sustainable and Privacy-Preserving Biometrics in Education

Supervisor: Tasmina Islam

Areas: Artificial Intelligence, Machine learning / Deep learning, Computer vision, Human-centred computing (human-computer interaction), Education

(Back to Scholarship Not Allocated)

Project Description

The growth of online and hybrid education has created new challenges for ensuring fairness and security in assessment. Traditional remote proctoring systems often depend on continuous webcam monitoring, raising concerns around privacy, accessibility, and environmental sustainability due to the energy and storage demands of video-heavy approaches.

This project aims to explore the design of sustainable and privacy-preserving biometric systems for education. The project will investigate how AI-driven biometric authentication methods can provide continuous but unobtrusive identity verification in online assessment while addressing three key issues:

Privacy-preserving authentication: developing methods that protect raw biometric data through on-device processing, secure template storage, or federated approaches.

Fairness and accessibility: auditing performance across diverse student groups and learning environments, ensuring transparency and inclusivity.

Ecological sustainability: benchmarking against conventional proctoring to quantify and reduce energy use, bandwidth, and carbon footprint.

The project will deliver prototype systems and evaluation frameworks that consider accuracy, fairness, privacy, and ecological impact together. The outcomes will contribute to the development of responsible, inclusive, and environmentally conscious online assessment technologies.

Foundation Models for Model-based Reinforcement Learning in Robotics

Supervisor: Matteo Leonetti

Areas: Machine learning / Deep learning, Robotics, Artificial Intelligence

(Back to Scholarship Not Allocated)

Project Description

Learning and effectively using world models is a currently unmet challenge in autonomous robotics. We will look into leveraging recent foundation models to build representation hierarchies, speed up policy search, and learn the world model itself.

Scaffolding Student Learning through GenAI in Cybersecurity Education

Supervisor: Hannah Cao / Ievgeniia Kuzminykh

Areas: Artificial Intelligence, Cybersecurity, Education

(Back to Scholarship Not Allocated)

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.

Self-Supervised Foundation Models for Video Panoptic Understanding

Supervisor: Luis C. Garcia Peraza Herrera

Areas: Artificial Intelligence, Machine learning / Deep learning, Computer vision

(Back to Scholarship Not Allocated)

Project Description

The development of foundation models has transformed AI, and video is the next frontier. Video Foundation Models (ViFMs) have shown great promise in high-level tasks like action recognition and text-based retrieval. However, there remains an unmet need for powerful, general-purpose representations that can support dense, pixel-level understanding tasks such as panoptic segmentation.

While self-supervised learning (SSL) has become the dominant paradigm for pre-training visual models without relying on labeled data, current SSL objectives are suboptimal for tasks requiring fine-grained, pixel-level understanding. Existing pretext tasks [1] focus on capturing motion and high-level semantics, making them unsuitable for panoptic segmentation.

Despite the importance of panoptic segmentation, there is no widely adopted self-supervised pre-training strategy specifically designed to learn representations tailored to this task in the video domain. This project aims to bridge this gap by developing a novel SSL objective that leverages the inherent structure and motion in unlabeled videos as a supervisory signal for learning panoptic representations.

We will design and train a new video foundation model using a novel self-supervised pre-training strategy, tailored to the specific requirements of panoptic segmentation. Our approach will focus on exploiting the structural information present in videos to learn rich, pixel-level representations that support complex scene understanding tasks.

References

[1] Zhao et al., "VideoPrism: A Foundational Visual Encoder for Video Understanding", ICML 2024.

AI-Driven Modelling of Immune Signalling Pathways in Solid Tumours

Supervisor: Dr Sophia Tsoka

Areas: Artificial Intelligence, Machine learning / Deep learning

(Back to Scholarship Not Allocated)

Project Description

Advances in artificial intelligence (AI) and machine learning (ML) now offer powerful opportunities to tackle complex biomedical problems. In cancer research, the challenge lies in integrating large, heterogeneous datasets (i.e. genomic, transcriptomic, immune receptor sequencing, and clinical data) into models that are both predictive and interpretable. Conventional approaches often provide descriptive analyses but fall short of generating robust frameworks that can forecast disease trajectories or treatment outcomes.

This project will develop novel computational methods to address these limitations, focusing on the analysis of solid tumours. It will employ deep learning—based integration approaches to combine multi-omic and clinical data, enhancing detection of immune cell states and signalling pathway activity. Network-based methods will be applied to reconstruct and analyse cell—cell communication and immune signalling interactions within the tumour microenvironment. In parallel, multiple instance learning will be used to model immune cell receptor repertoires and link them to tumour progression and therapy response. Together, these frameworks will enable the construction of predictive, digital twin—like models of immune behaviour in cancer.

The biological motivation for this work lies in the pressing need to understand immune mechanisms underlying tumour development and treatment response. Solid tumours such as melanoma, breast, lung, and colorectal cancer display striking variability in their clinical course and in patient responses to immunotherapy. While immune checkpoint inhibitors and related treatments have revolutionised oncology, many patients derive little or no long-term benefit. To improve prognosis and enable personalised therapy, it is critical to uncover the immune cell populations, signalling pathways and network interactions that drive these differences.

The project's outcomes will include:

- Computational tools for representing and analysing tumour—immune signalling networks.
- Predictive models of therapy response with translational potential.
- Open-access resources supporting both computational and biomedical research communities.

Through advanced computational innovation and alignment with clinical impact, this project bridges data science and cancer immunology. Its results will not only advance methodological frontiers in AI/ML but also deliver actionable insights into tumour—immune interactions, with potential to transform cancer diagnosis, prognosis, and treatment.

References

1. da Costa Avelar PH, Laddach R, Karagiannis SN, Wu M, Tsoka S (2023). Multi-omic Data Integration and Feature Selection for Survival-Based Patient Stratification via Supervised Concrete Autoencoders. In: Nicosia, G., et al. Machine Learning, Optimization, and Data Science. LOD 2022. Lecture Notes in Computer Science, vol 13811. Springer, Cham. https://doi.org/10.1007/978-3-031-25891-6_5

2. Crescioli S, et al. (2023). B cell profiles, antibody repertoire and reactivity reveal dysregulated responses with autoimmune features in melanoma. Nat Commun. 14(1):3378. https://doi.org/10.1038/s41467-023-39042-y

3. Amiri Souri E, Chenoweth A, Cheung A, Karagiannis SN, Tsoka S (2021). Cancer Grade Model: a multi-gene machine learning-based risk classification for improving prognosis in breast cancer. Br J Cancer 125, 748—758. https://doi.org/10.1038/s41416-021-01455-1

4. Chen Y, Liu S, Papageorgiou LG, Theofilatos K, Tsoka S (2023). Optimisation Models for Pathway Activity Inference in Cancer. Cancers. 2023; 15(6):1787. https://doi.org/10.3390/cancers15061787

5. Liapis GI, Tsoka S, Papageorgiou LG (2025). Optimisation-Based Feature Selection for Regression Neural Networks Towards Explainability. Machine Learning and Knowledge Extraction, 7(2):33. https://doi.org/10.3390/make7020033