Machine Learning / Deep Learning 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.

Game theory, learning and mechanism design in cryptoeconomic systems

Supervisor: Stefanos Leonardos

Areas: Machine learning / Deep learning, Cryptoeconomics, Game theory

(Back to Scholarship Not Allocated)

Project Description

This project is for students interested in exploring the fast-growing field of cryptoeconomics through the lens of game theory. The focus is on modeling and analyzing blockchain-based systems to understand the incentives that drive participants and the mechanisms that shape outcomes. Topics include transaction fee mechanisms (TFMs), miner extractable value (MEV), automated market makers (AMMs), transaction censorship, and many more. Students will investigate how cryptoeconomic mechanisms evolve, how strategic behavior impacts blockchain performance, and how new mechanisms can be designed for greater efficiency and fairness. The project combines tools from game theory, economics, computer science, and dynamical systems, with opportunities to build models, run simulations, and connect theory with real-world blockchain applications.

References

1. Buterin, V, Reijsbergen, D, Leonardos, S, Piliouras, G. Incentives in Ethereum's hybrid Casper protocol. Int J Network Mgmt. 2020; 30:e2098. https://doi.org/10.1002/nem.2098
2. Leonardos, S., Reijsbergen, D., Monnot, B., and Piliouras, G., "Optimality Despite Chaos in Fee Markets", arXiv e-prints, 2022. doi:10.48550/arXiv.2212.07175, (2025).
3. Performative Market Making, C Kleitsikas, S Leonardos, C Ventre, arXiv preprint arXiv:2508.04344.
4. MEV Sharing with Dynamic Extraction Rates, P Braga, G Chionas, P Krysta, S Leonardos, G Piliouras, C Ventre, Proceedings of the Workshop on Decentralized Finance and Security, 1-10, (2024).
5. W. Wu, T. Thiery, S. Leonardos, C. Ventre. Strategic Bidding Wars in On-chain Auctions. IEEE ICBC 2024, https:/arxiv.org/abs/2312.14510.
6. Leonardos, S., Monnot, B., Reijsbergen, D., Skoulakis, E., and Piliouras, G. (2021). Dynamical analysis of the EIP-1559 Ethereum fee market. In Proceedings of the 3rd ACM Conference on Advances in Financial Technologies (AFT '21). Association for Computing Machinery, New York, NY, USA, 114—126. https://doi.org/10.1145/3479722.3480993.

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

Software sustainability analysis and improvement

Supervisor: Kevin Lano

Areas: Machine learning / Deep learning, Systems (software engineering, programming), Foundations of computing (algorithms, computational complexity)

(Back to Scholarship Not Allocated)

Project Description

The project would consider techniques for analysing software sustainability (in the sense of energy use and energy efficiency) 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.

Energy-efficiency improvement of machine learning systems is particularly important and could be the focus of the research. Equally, energy-efficiency improvement of mobile apps is another possible focus.

There is the potential for industrial collaboration in this area.

References

(Lano et al., 2024a) K. Lano et al., "Software modelling for sustainable software engineering", STAF 2024.
(Lano et al., 2024b) K. Lano et al., "Design Patterns for Software Sustainability", PLoP 2024
(Lano et al, 2025) K. Lano et al, "Sustainable Software Re-engineering", ECMFA 2025.

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

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.

Communication, Information, and Robustness in Trading Networks

Supervisor: Edwin Lock / Carmine Ventre

Areas: Machine learning / Deep learning, Foundations of computing (algorithms, computational complexity), Algorithmic Game Theory

(Back to Scholarship Not Allocated)

Project Description

Many markets function without a central auctioneer, instead relying on decentralised interactions between participants in environments characterised by bilateral negotiation and limited information exchange. This project will investigate the computational foundations of such decentralised trading. It will combine tools from algorithmic game theory, decentralised computing, and machine learning to explain how stable and equitable outcomes can arise, or fail to arise, when agents interact strategically in different informational environments.

This project focuses on how the amount and type of information exchanged between agents shapes outcomes in decentralised markets. Examples of such markets include financial markets and trading mechanisms implemented on the blockchain. In these settings, communication can be costly (e.g., blockchain transaction fees) or sensitive (e.g., a reluctance to reveal strategically valuable data). The project will investigate trade-offs between communication costs and the attainability of stable outcomes, as well as how incomplete information and strategic misrepresentation affect equilibrium behaviour. It will draw on tools from communication complexity and mechanism design, complemented by multi-agent reinforcement learning to simulate negotiation protocols under different informational constraints.

The student will gain expertise in algorithmic game theory, complexity analysis, market design, and machine learning. They will join a cross-disciplinary research environment at the intersection of computer science and economics, with potential opportunities for collaboration in both academia and industry (e.g. financial markets, commodity markets, and digital marketplaces). The project provides training in both rigorous mathematical analysis and computational experimentation, equipping the student with a broad skillset relevant to academia, industry, and policy.

Lifelong and Culturally-Aware Model Editing: Towards Scalable, Context-Sensitive Knowledge Updates in Multilingual Language Models

Supervisor: Helen Yannakoudakis

Areas: Machine learning / Deep learning, Natural Language Processing

(Back to Scholarship Not Allocated)

Project Description

Large language models are increasingly deployed in dynamic, multilingual settings where facts, APIs, and public knowledge shift frequently. Existing model editing techniques such as ROME, MEMIT, and MEND have shown success in making localised factual corrections, but they typically assume that edits are isolated, language-neutral, and structurally simple. These assumptions break down in real-world applications, where edits must accumulate over time and reflect differing cultural or linguistic realities. In practice, a factual update that is accurate and appropriate in English may require significant adaptation, or even a different framing entirely, when expressed in Arabic, Hindi, or Swahili. Such differences are not mere translation artifacts but stem from deeper issues of cultural context and knowledge representation.

This PhD project proposes a new framework for lifelong and culturally-aware model editing that supports both the scalability of long-horizon updates and the sensitivity required for cross-lingual propagation. At its core is a dual-phase editing system: rapid, lightweight interventions are used for immediate local updates, while slower consolidation phases ensure long-term stability through replay-based regularisation and continual learning techniques. The project is grounded in two central challenges: first, how to ensure that models remain stable and composable under long sequences of factual edits, without suffering interference, drift, or catastrophic forgetting; and second, how to propagate a single factual update across multiple languages in a way that preserves semantic intent while allowing for appropriate cultural and contextual variation, particularly in low-resource or typologically distant languages.

To address these challenges, the framework introduces editing in a shared causal latent space designed to align internal representations across languages, enabling coherent generalisation without assuming uniformity. At the same time, it permits language-specific or culture-specific adaptation when edits carry different implications across linguistic contexts. By integrating insights from model editing, multilingual NLP, and lifelong learning, this work aims to advance the reliability, adaptability, and cultural robustness of foundation models operating in a global context.

References

[1] Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? EMNLP, 2023.
[2] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 1(1):12, 2021.
[3] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In International conference on machine learning. PMLR, 2020.
[4] Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Vlkeb: A large vision-language model knowledge editing benchmark. NeurIPS, 2024.
[5] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521—3526, 2017.
[6] Jiaang Li, Quan Wang, Zhongnan Wang, Yongdong Zhang, and Zhendong Mao. Enhance lifelong model editing with continuous data-adapter association. arXiv e-prints, pages arXiv—2408, 2024.
[7] Zhizhong Li and Derek Hoiem. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12):2935—2947, 2017.
[8] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. NeurIPS, 2017.
[9] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. Advances in neural information processing systems, 35:17359—17372, 2022.
[10] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. ICLR, 2023.
[11] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. ICLR.
[12] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. ICLR. PMLR, 2022.
[13] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. ICLR.
[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. NeurIPS, 2022.
[15] Jonas Pfeiffer, Aishwarya Kamath, Andreas Ruckle, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. EACL, 2021.
[16] Jonas Pfeiffer, Ivan Vuli´c, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. EMNLP, 2020.
[17] Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. Cross-lingual knowledge editing in large language models. ACL, 2024.
[18] Mengqi Zhang, Xiaotian Ye, Qiang Liu, Shu Wu, Pengjie Ren, and Zhumin Chen. Uncovering overfitting in large language model editing. ICLR, 2025.
[19] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. NeurIPS.

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

Language-aware alignment and cross-lingual safeguards in LLMs

Supervisor: Oana Cocarascu

Areas: Natural Language Processing, Machine learning / Deep learning

(Back to Scholarship Not Allocated)

Project Description

Large language models (LLMs) have shown disparities in their responses within and across languages. Adversaries can exploit inconsistent behaviour in low-resource languages to bypass safety measures, spread misinformation or generate harmful content. With current evaluations and benchmarks primarily focusing on English, there is a need for language-aware alignment and cross-lingual safeguards in LLMs.

This PhD project aims to conduct a systematic evaluation of the behaviour of LLMs in multilingual settings and use interpretability techniques to examine the mechanisms behind unsafe generation to improve safety across languages. Concretely, it will focus on low-resource languages, dialectal variations within high-resource languages, and code-mixed languages. Through extensive evaluations across multiple languages and variations thereof, the project will provide a better understanding of model behaviour on safety-sensitive inputs across languages. To uncover the mechanisms leading to unsafe generation in LLMs across diverse languages, we will identify latent regions tied to safety as these can vary between languages or dialects, and perturb components to determine latent risks across diverse linguistic inputs. By analysing LLM responses to prompts in multiple languages, this work will uncover language-specific risks and will provide novel insights into the interaction between language and unsafe behaviour.