PhD projects (Human-centred computing)

- Urban AI Shields: GeoAI and Multimodal Data for Predicting Extreme Weather Health Impacts on Vulnerable Urban Populations
- Semantic Visual Analytics for Exploring the Evolution of Folklore across Cultures and Media
- Human-Centred Visual Analytics for Interrogating LLM-Generated Knowledge
- Evolving Cybercrime: Improving Threat Recognition and Response
- Designing Against Intimate Partner Violence in AI Companions
- Immersive Visual Analytics for Exploration, Debugging and Understanding of Quantum Algorithms and Machine Learning Models
- Visualisation for Quantum Computing
- Investigating Bertin's Retinal Variables through Quantum Cognition
- Visual Analytics for Environmental Forecasting Across Space, Time and Uncertainty
- Human-Centred AI for Trustworthy Environmental Forecasting and Decision-Making
- Civic and Responsible AI
- Adaptive Human-AI Collaboration for Reasoning and Learning
- Human-Centred and Adaptive Visualization for an Ageing Population
- Designing visual analytics systems that capture learning, expertise and ageing as interacting factors for optimal decisions
- Modelling Behavioural Change for Adaptive Biometrics
- Enabling Accessible Remote Communication Environments
- Embodied Approaches to Assistive Technology
- Privacy in AI Companion / Assistant Technologies
- Scalable Oversight of AI Agents
- Self-Evolving Multi-Agent LLM Systems: Collaborative Training and Recursive Self-Improvement
- Scaffolding Student Learning through GenAI in Cybersecurity Education
- Validation of Cyber Threat Intelligence Leveraging Large Language Models

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

Supervisor: Yijing Li

Project Description

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

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

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

Supervisor: Alfie Abdul-Rahman

Project Description

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

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

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

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

References

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

Human-Centred Visual Analytics for Interrogating LLM-Generated Knowledge

Supervisor: Alfie Abdul-Rahman

Project Description

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

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

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

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

References

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

Evolving Cybercrime: Improving Threat Recognition and Response

Supervisor: Tasmina Islam / Kovila Coopamootoo

Project Description

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

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

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

References

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

Designing Against Intimate Partner Violence in AI Companions

Supervisor: Dr Ruba Abu-Salma and Dr Martim Brandao

Project Description

This project examines how sycophancy and aggression in AI companions are understood by users and by intimate partners who may be affected without using these systems themselves. It investigates how conversational patterns that affirm users' beliefs, simulate intimacy, or encourage harmful behavior may shape relationships and create risks that existing AI safety approaches overlook. Grounded in feminist standpoint epistemology, the research centers the knowledge of intimate partner violence survivors while attending to cultural differences and inequalities within each setting. The research will proceed through three connected phases. Interviews with users of AI companions and general-purpose models used for romantic conversations will explore interpretations of agreement, emotional validation, and influence. Workshops and interviews with survivors, recruited through trusted community organizations, will examine their understandings of these behaviors and the protections they consider meaningful. Findings will then inform co-design workshops with new user and non-user cohorts to develop a culturally differentiated corpus of harmful or contested interactions. Users, non-users, and AI safety researchers will collaboratively develop design recommendations and a working prototype, refined through iterative feedback. The project will contribute empirical accounts of relational harm, culturally grounded resources for safety evaluation, and practical design interventions addressing both user and non-user safety. Throughout, participation will prioritize informed consent, control over disclosure, and recognition of participants' knowledge and labor.

References

Position: Responsible AI for AI Companions Must Actively Combat Violence Toward Intimate Partners: https://openreview.net/forum?id=s6nxLbamu4

Immersive Visual Analytics for Exploration, Debugging and Understanding of Quantum Algorithms and Machine Learning Models

Supervisor: Rita Borgo / Mohammad Reza Mousavi

Project Description

This PhD project investigates how immersive and interactive visual analytics can support the development and understanding of quantum computing systems. As quantum algorithms and machine learning models grow in complexity, their behaviour becomes increasingly difficult to interpret with traditional tools. The project aims to bridge this gap by designing novel visual and immersive environments that make the inner workings of quantum programs more accessible and comprehensible.

The research will explore techniques for visualising state space dynamics and quantum evolution, combining these with VR/AR interfaces to support spatial exploration and intuitive interaction. Such tools have the potential to transform how researchers debug, train, and interpret quantum systems, improving transparency, interpretability, and trust in quantum software.

Visualisation for Quantum Computing

Supervisor: Rita Borgo / Mohammad Reza Mousavi

Project Description

Quantum computing represents a profound shift in how information is processed and analysed, but its development is constrained by the difficulty of understanding and debugging quantum programs. Quantum algorithms and quantum machine learning (QML) models operate in high-dimensional Hilbert spaces, with entangled and probabilistic states that are challenging to interpret using classical tools. This creates a critical need for human-centred approaches that make quantum processes more transparent, interpretable, and usable.

This PhD project will investigate the design of novel visualisation and visual analytics methods to support quantum software development and research. Key directions include techniques for visualising quantum state evolution, entanglement, and interference patterns; approaches to debugging and verifying quantum programs through visual diagnostics.

Investigating Bertin's Retinal Variables through Quantum Cognition

Supervisor: Rita Borgo / Andrew Meso

Project Description

This PhD project explores a novel application of quantum cognitive theory to the study of Bertin's retinal variables—fundamental visual encodings such as colour, shape, size, and position. While traditional models treat these variables as independent channels processed in isolation, emerging evidence suggests that perception may instead involve non-classical, quantum-like dynamics. This research aims to investigate whether interpretations of visual encodings exhibit phenomena such as superposition, interference, and contextuality.

Visual Analytics for Environmental Forecasting Across Space, Time and Uncertainty

Supervisor: Rita Borgo / Emma Tebbs

Project Description

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

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

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

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

Supervisor: Rita Borgo / Emma Tebbs

Project Description

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

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

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

Civic and Responsible AI

Supervisor: Martim Brandao

Project Description

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

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

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

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

Adaptive Human-AI Collaboration for Reasoning and Learning

Supervisor: Tasmina Isalm

Project Description

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

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

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

References

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

Human-Centred and Adaptive Visualization for an Ageing Population

Supervisor: Rita Borgo / Andrew Meso / Peter Baudains

Project Description

As populations age, visualization systems must support users with increasingly diverse perceptual, cognitive, and interaction capabilities. This PhD will investigate how computational models of human perception and behaviour can inform the design of visualizations and visual analytics systems that remain effective across the lifespan.

The project will study how established factors associated with ageing, including changes in perception, cognition and motor behaviour affect the interpretation of complex and dynamic visualisations. Controlled experiments using behavioural measures and eye tracking will be carried out to measure participant sensitivity and individual differences across the lifespan for a set of spatiotemporal and design parameters. The results will be used to develop statistical and dynamic computational modelling to characterise and predict age-related differences in visualization performance.

These complementary empirical and modelling insights will inform the design and implementation of adaptive visualization techniques that can modify visual encodings, complexity, interaction, or information presentation according to users' capabilities and task requirements. Outcomes will be evaluated across different age groups and visual analytic tasks. Expected outcomes include computational models of visualization use across the lifespan, novel adaptive and age-inclusive visualization techniques, and evidence-based design guidelines for accessible visual analytics.

The project sits at the intersection of visualization, visual analytics, HCI, human-centred AI, accessibility, and computational modelling.

Designing visual analytics systems that capture learning, expertise and ageing as interacting factors for optimal decisions

Supervisor: Rita Borgo / Andrew Meso / Peter Baudains

Project Description

An increasing prevalence of AI and algorithmic decision support systems for the control and management of socio-technical systems, such as transport networks and the power grid, continue to rely on visual interfaces to keep human operators 'in the loop'. Rather than fully automated decisions, such systems instead present scenarios and predicted outcomes so that operators retain agency and responsibility by applying human judgment to the proposed algorithmic options. This judgment can bring significant expertise to validate solutions and identify issues not considered by the algorithm. For example, using their experience, operators may be able to quickly identify unworkable proposals due to organisational constraints or even ethical reasons.

This project will consider the design of visualisation and visual analytics systems that support integration of human expertise within the context of a diversity in the levels of operator experience within teams and an aging workforce with impacts on the most experienced colleagues. Expertise may be built up over a long period of time, sometimes even decades. This brings two related challenges: 1. How to design systems that best support expertise in the context of aging, for example by accounting for established cognitive aging effects? 2. How can such systems concurrently use visualisation to better support newer or novice users, who may not have as much experience to bring to bear on expert-informed judgments?

Through empirical experiments and user studies with both expert and novice users complemented by normative population modelling of the results, this work will provide design guidelines and prototypes for visual analytics decision support systems that promote and support the integration of expertise in decision making across the lifespan.

Modelling Behavioural Change for Adaptive Biometrics

Supervisor: Tasmina Islam

Project Description

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

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

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

References

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

Enabling Accessible Remote Communication Environments

Supervisor: Timothy Neate

Project Description

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

References

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

Embodied Approaches to Assistive Technology

Supervisor: Timothy Neate

Project Description

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

References

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

Privacy in AI Companion / Assistant Technologies

Supervisor: Kovila Coopamootoo

Project Description

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

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

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

Scalable Oversight of AI Agents

Supervisor: Yali Du

Project Description

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

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

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

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

References

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

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

Supervisor: Yali Du

Project Description

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

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

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

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

References

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

Scaffolding Student Learning through GenAI in Cybersecurity Education

Supervisor: Hannah Cao / Ievgeniia Kuzminykh

Project Description

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

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

References

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

Validation of Cyber Threat Intelligence Leveraging Large Language Models

Supervisor: Hannah Cao

Project Description

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

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

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

References

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

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

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