PhD projects in the Department of Informatics, AY 25-26 — Artificial Intelligence (symbolic AI, logic, etc.)

The PhD projects listed below will be considered for 2025/26 studentships available in the Department of Informatics to start on 1 October 2025 or later during the 2025/26 academic year.

Please note that this list is not exhaustive and potential applicants can alternatively identify and contact appropriate supervisors outlining their background and research interests or proposing their own project ideas.

Each project is designated for a single student, meaning it can only be assigned to one successful applicant. Some projects come with allocated studentships, while others are eligible for "unallocated" studentships. Applicants who apply for projects with allocated studentships and are selected will be offered a full studentship. In the project list, these are marked as "studentship allocated." Applicants chosen for other projects will compete for the unallocated studentships.

We welcome applications from students who have secured, or are applying for, or plan to apply for other funding (within other studentships internal to the university or external schemes) and from self-funded students. See also this list of funding opportunities available at King's for post-graduate research in Computer Science.

PhD projects

- AI and NLP for Multilingual Code-Switching in Education (studentship allocated)
- Leveraging Generative AI for Creativity Education (studentship allocated)
- Improving active learning strategies for limited annotation budgets (studentship allocated)
- Character-Centric Systems for Multimodal Story Generation (studentship allocated)
- Towards Robust Reasoning of Large Language Models (studentship allocated)
- Allowing autonomous robots to continually learn, generalize, and improve from their experiences (studentship allocated)
- Game-theoretic models in cryptoeconomics: incentives, mechanism design and blockchain dynamics
- Game-theoretic models in multi-agent systems: emergent behaviours, critical phase transactions and learning dynamics
- Sustainable Delivery Logistics with Drones and Cargo Bikes
- Goal-based explanations for autonomous systems and robots
- Leveraging Language Models for Contextual Vulnerability Identification
- Advanced Modelling on Multimodal Urban Geospatial Data Fusion - Case Studies for UK Cities
- Exploring Interactive Multi-Dimensional Approaches of Delivery of Communication in Patient Scenarios in Oral Health Education
- Software sustainability analysis and improvement
- Safe Reinforcement Learning from Human Feedback
- Multi-agent Cooperation with RL and LLMs
- Assessing the value of evidence with argument-driven credal networks
- Investigating LLM-based Generative AI Applications in Cybersecurity
- Argument mining
- Multilingual argument mining
- AI in finance
- Adaptation and effective communication in collaborative physically Assistive Tasks
- Explaining robotic planning decision points along execution
- Enhancing Safety in Robotics by Tackling Blind-Spots and Bias in AI Models
- Agents powered by foundation models
- Causal Explanations for Sequential Decision Making
- Reliable Learning for Safe Autonomy with Conformal Prediction
- Planning and Reinforcement Learning for versatile autonomous robots
- Designing and Developing a framework for responsible security and privacy practices for GenAI Tools
- Implementing Differential Privacy in Neural Networks to Enhance Data Security and Anonymization
- Discovering the Secrets of Random Neural Networks - Training by Pruning

AI and NLP for Multilingual Code-Switching in Education

Supervisor: Zheng Yuan

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

Project Description

The rapid growth of multilingualism has led to an increased prevalence of code-switching (CSW) -- the practice of alternating between two or more languages within a single conversation or utterance. Despite its common usage in multilingual communication, current Natural Language Processing (NLP) technologies struggle to handle CSW effectively, particularly in educational contexts. This project aims to address this gap by developing advanced NLP technologies and educational AI systems specifically designed to support multilingual CSW environments. The goal is to create a personalised, inclusive, and engaging AI-powered tutoring system that adapts to the unique linguistic needs of learners. This project will focus on one or more of the following key areas: 1) Development of NLP models that can accurately process and analyse CSW data, distinguishing code-switching from grammatical errors; 2) Creation of an Intelligent Tutoring System (ITS) that provides personalised feedback and assessment tailored to the needs of multilingual learners; 3) Leveraging multilingual Large Language Models (LLMs) to enhance the capabilities of AI in educational settings, particularly in low-resource languages; and 4) Evaluating the impact of educational AI systems in real-world settings, assessing improvements in learning outcomes, learner engagement, and satisfaction.

References

- LLM-based Code-Switched Text Generation for Grammatical Error Correction. Tom Potter and Zheng Yuan. EMNLP 2024.
- Prompting open-source and commercial language models for grammatical error correction of English learner text. Christopher Davis, Andrew Caines, Øistein E. Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei and Paula Buttery. ACL 2024 Findings.
- Grammatical Error Correction for Code-Switched Sentences by Learners of English. Kelvin Chan, Christopher Bryant, Li Nguyen, Andrew Caines and Zheng Yuan. LREC-COLING 2024.
- Grammatical Error Correction. Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng and Ted Briscoe. Computational Linguistics; https://doi.org/10.1162/coli_a_00478
- Building Educational Technologies for Code-Switching: Current Practices, Difficulties and Future Directions. Li Nguyen, Zheng Yuan and Graham Seed. Languages; https://doi.org/10.3390/languages7030220

Leveraging Generative AI for Creativity Education

Supervisor: Zheng Yuan

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Computer vision, Machine learning / Deep learning, Human-centred computing (human-computer interaction), Natural Language Processing

Project Description

Creativity is a crucial skill in today's world, driving innovation, problem-solving, and cultural expression. However, teaching and assessing creativity -- especially in fields like creative writing and visual arts -- pose significant challenges due to the subjective nature of creative outputs. The rise of Large Language Models (LLMs) and Generative AI provides new opportunities for enhancing creativity education by generating personalised, adaptive feedback and supporting learners in improving their creative skills across multiple modalities, such as writing and drawing. This project will explore the use of multimodal LLMs and Generative AI to enhance creativity education, offering new ways to assess creativity and helping learners across disciplines such as creative writing and digital arts. By leveraging the capabilities of multimodal models, this research will investigate how AI can support, nurture, and assess creativity in a personalised and scalable manner. Research Questions: 1) How can LLMs and Generative AI effectively assess creativity in different forms, such as written stories, poems, or drawings? 2) What are the most effective ways for AI systems to provide feedback that nurtures creativity, without stifling originality? 3) How can multimodal AI systems enhance cross-disciplinary creative education (e.g. combining writing and drawing) to create richer, more engaging learning experiences? 4) What metrics and frameworks can be developed to evaluate the success of AI-generated feedback and creativity assessment systems?

Improving active learning strategies for limited annotation budgets

Supervisor: Luis C. Garcia Peraza Herrera

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, Computer vision

Project Description

In machine learning, determining the subset of data points (e.g. images, videos) for annotation emerges as a critical decision-making process. The selected data points carry the responsibility of providing a representative snapshot of the diverse scenarios anticipated during real-world testing. Despite the multitude of proposed strategies for data point selection, an enduring observation persists, suggesting that random selection, especially in low-budget scenarios, often proves to be an optimal approach. The overarching objective of this project is to propel active learning strategies tailored specifically for situations characterized by highly limited annotation budgets. This pursuit is particularly relevant in fields with stringent budget constraints, such as medicine.

References

https://visurg.ai/join

Character-Centric Systems for Multimodal Story Generation

Supervisor: Lin Gui/Yulan He

Areas: Machine learning / Deep learning, Artificial Intelligence (symbolic AI, logic, etc.), Natural Language Processing

Project Description

The primary goal of this project is to design and develop a character-centric multimodal system capable of generating rich, coherent narratives from multimodal inputs. This system will focus on story generation based on different types of data—such as images or audio—and could be applied in a variety of settings, including museums, medical diagnostics, or educational explanations. Specifically, the system would be able to generate detailed descriptions, historical accounts, or explanations from a given image or set of multimodal data. The research problem consists of several interconnected challenges that need to be addressed to achieve the goal: multimodal input interpretation, text generation based on input data, character-centric storytelling, and cross-domain adaptability. By focusing on a character-driven approach and cross-domain adaptability, the proposed system will not only engage users but also deliver accurate, contextually relevant content based on diverse input types. This system holds great potential for enhancing user experience in numerous real-world applications, driving innovation in AI-based storytelling and explanation systems.

References

1. Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, Antoine Bosselut: PeaCoK: Persona Commonsense Knowledge for Consistent and Engaging Narratives. ACL (1) 2023: 6569-6591.

2. Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, Yulia Tsvetkov: Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. ACL (1) 2023: 13960-13980.

3. Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, Ziwei Liu: Octopus: Embodied VisionLanguage Programmer from Environmental Feedback. CoRR abs/2310.08588 (2023).

4. Yujie Wang, Hu Zhang, Jiye Liang, Ru Li: Dynamic Heterogeneous-Graph Reasoning with Language Models and Knowledge Representation Learning for Commonsense Question Answering. ACL (1) 2023: 14048-14063.

5. Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, Lin Gui: NarrativePlay: Interactive Narrative Understanding. CoRR abs/2310.01459 (2023).

6. Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, Gao Huang: Avalon's Game of Thoughts: Battle Against Deception through Recursive Contemplation. CoRR abs/2310.01320 (2023).

7. Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, Yang Liu: Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. CoRR abs/2309.04658 (2023).

8. Lixing Zhu, Runcong Zhao, Lin Gui, Yulan He: Are NLP Models Good at Tracing Thoughts: An Overview of Narrative Understanding. CoRR abs/2310.18783 (2023).

Towards Robust Reasoning of Large Language Models

Supervisor: Yulan He

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, Natural Language Processing

Project Description

Context Reasoning is a core aspect of human intelligence, essential for tasks such as critical thinking, evaluation and making decisions. With the advancements of large language models (LLMs), we have witnessed their impressive performance in various natural language processing tasks that require reasoning. For an intelligent system to be effective, it must thoroughly analyse key information within a given context and provide accurate responses by leveraging its internal knowledge and external resources. This is a complex process as LLMs need to stay current with new information, remain robust in noisy contexts, and be capable of utilising external tools for validation when necessary.

Project: Despite advancements in the reasoning capabilities of LLMs, there remains uncertainty regarding the extent to which LLMs can reason beyond memorisation. Recent empirical studies have highlighted their susceptibility to challenges posed by noisy contexts, new information, and novel tasks. Therefore, our goal is to create a robust reasoning framework that enables LLMs to reason effectively when presented with new and unfamiliar inputs. To achieve this, example tasks include:

- Enhancing reasoning through tool augmentation based on a neuro-symbolic approach. LLMs can improve their reasoning by leveraging neuro-symbolic methods with the help of external interpreters, particularly in more complex tasks.
- Facilitating model adaptation to reason with the most recent knowledge. This involves model editing and fine-tuning LLMs with new information while ensuring they retain their reasoning abilities for previously encountered tasks.
- Encouraging collaboration among multiple LLM agents to support reasoning across diverse domains. When faced with an input from an unfamiliar domain, integrating knowledge from multiple trained LLM agents based on its relevance to the specific input could enhance reasoning performance.

References

- Mirzadeh I, Alizadeh K, Shahrokhi H, Tuzel O, Bengio S, Farajtabar M. 2024. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. [pdf]
- Jie H, Kevin Chen-Chuan C. 2023. Towards Reasoning in Large Language Models: A Survey. [pdf]
- Collin B, Haotian Y, Dan K, Jacob S. 2022. Discovering Latent Knowledge In Language Models Without Supervision. [pdf]
- Almog G, Elad V, Colin R, Noam S, Yoav K, Leshem C. 2023. Knowledge is a Region in Weight Space for Fine-tuned Language Models. [pdf]
- Luyu G, Aman M, Shuyan Z, Uri A, Pengfei L, Yiming Y, Jamie C, Graham Ng. 2023. PAL: Program-aided Language Models. [pdf]
- Marco F, Florian W, Luca Z, Alessandro A, Emanuele R, Stefano S, Bernhard S, Francesco L. 2023. Leveraging sparse and shared feature activations for disentangled representation learning. [pdf]
- Jonas P. Sebastian R. Ivan V.. Edoardo M. P∗.2023. Modular Deep Learning. [pdf]

Allowing autonomous robots to continually learn, generalize, and improve from their experiences

Supervisor: Dr. Khen Elimelech

Areas: Robotics, Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning

Project Description

To perform autonomously tasks such as object rearrangement, assembly, manipulation, and navigation, robots must be able to plan their actions over long horizons. Such planning is usually computationally challenging to perform in real time, especially considering complex robots and task specifications, or large and uncertain planning domains, with many irrelevant objects and distractions. One intuitive approach to support autonomous robots in this challenge is by allowing them to learn to continually improve their planning capabilities over time, based on their experience. This general approach should enable us to build long-lived, multi-purpose robots with human-like versatility and common sense, rather than highly specialized machines.

Unfortunately, despite recent advancement in Machine Learning and "Learning from Demonstrations," existing learning approaches are not suitable for this objective, as these require numerous annotated demonstrations, rendering them unsuitable for online, autonomous learning.

To this end, our recent work introduced a novel algorithmic framework for automatic learning of "planning strategies" by abstracting successful planning experiences. This framework allows a robot to automatically and continually make generalizable conclusions from individual experiences, which can later be adapted for and reused in new contexts, to accelerate the solution of new planning problems—just like humans do, but without human intervention!

Initial results demonstrated the potential of this approach to significantly impact the field of AI-enabled robotics. To achieve that, this project seeks to extend this initial effort in various directions, including: application and adaptation to new platforms, planning domains, and task-types; application to multi-robot and human-robot collaborative systems; integration with (statistical) Machine Learning and Computer Vision techniques, Control, knowledge graphs and other components in the autonomy stack; improving utility and computational tractability through algorithmic development; and improving trustworthiness through formal analysis.

The work on this project is diverse and contains theoretical, computational. and experiential aspects. Students are expected to conduct research, publish papers, develop and release open-source code, and work with physical robots. You will have access to state-of-the-art hardware and resources, and excellent mentorship. Potentially, successful students will have access to collaboration and internship opportunities with industry leaders, such as NASA Robotics, Amazon Robotics and Bosche.

While prior research experience in robotics is recommended, it is not mandatory. Excellent candidates with background in robotics, AI, computer science, algorithms, applied mathematics, engineering, or other relevant background are welcome to apply.

References

[1] Accelerating Long-Horizon Planning with Affordance-Directed Dynamic Grounding of Abstract Strategies, by Khen Elimelech, Zachary Kingston, Wil Thomason, Moshe Y. Vardi, and Lydia E. Kavraki, in IEEE International Conference on Robotics and Automation (ICRA), May 2024.

[2] Extracting generalizable skills from a single plan execution using abstraction-critical state detection, by Khen Elimelech, Lydia E. Kavraki, and Moshe Y. Vardi, in IEEE International Conference on Robotics and Automation (ICRA), May 2023.

[3] Principles of Robot Motion: Theory, Algorithms, and Implementations, by Howie Choset, Kevin M. Lynch, Seth Hutchinson, George A. Kantor, Wolfram Burgard, Lydia E. Kavraki and Sebastian Thrun, The MIT Press, 2005.

Game-theoretic models in cryptoeconomics: incentives, mechanism design and blockchain dynamics

Supervisor: Dr. Stefanos Leonardos

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, Foundations of computing (algorithms, computational complexity), Game theory

Project Description

This project is aimed for students who are interested in advancing cutting-edge research at the intersection of game theory and cryptoeconomics. The project will focus on modelling and analyzing blockchain-enabled economies through a game-theoretic lens. Special focus will be placed on transaction fee mechanisms (TFMs), miner extractable value (MEV), proposer-builder separation (PBS) in Ethereum block creation, MEV-boost auctions, dynamics of automated market makers (AMMs), transaction censorship, attacks in decentralized exchanges, and related phenomena. The study will explore cryptoeconomic mechanisms, dissect participants' incentives, and designing mechanisms to optimize blockchain performance. Due to the dynamic nature of these systems, the project will employ elements from algorithmic game theory and dynamical systems, alongside standard tools from economics, computer science, and machine learning. Successful candidates will develop game-theoretic models, conduct rigorous mathematical analyses, and run simulations to validate theoretical predictions in real-world applications, bridging the gap between academia and industry.

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.
3. W. Wu, T. Thiery, S. Leonardos, C. Ventre. Strategic Bidding Wars in On-chain Auctions. IEEE ICBC 2024, https://arxiv.org/abs/2312.14510.
4. Leonardos, S, Reijsbergen, D, Piliouras, G. Weighted voting on the blockchain: Improving consensus in proof of stake protocols. Int J Network Mgmt. 2020; 30:e2093. https://doi.org/10.1002/nem.2093
5. Leonardos, N., Leonardos, S., Piliouras, G. (2020). Oceanic Games: Centralization Risks and Incentives in Blockchain Mining. In: Pardalos, P., Kotsireas, I., Guo, Y., Knottenbelt, W. (eds) Mathematical Research for Blockchain Economy. Springer Proceedings in Business and Economics. Springer, Cham. https://doi.org/10.1007/978-3-030-37110-4_13
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
7. D. Reijsbergen, S. Sridhar, B. Monnot, S. Leonardos, S. Skoulakis and G. Piliouras, "Transaction Fees on a Honeymoon: Ethereum's EIP-1559 One Month Later," 2021 IEEE International Conference on Blockchain (Blockchain), Melbourne, Australia, 2021, pp. 196-204, doi: 10.1109/Blockchain53845.2021.00034.
8. Koki, C., Leonardos, S., and Piliouras, G. (2022). Exploring the predictability of cryptocurrencies via Bayesian hidden Markov models, Research in International Business and Finance, Volume 59, 101554, doi: 10.1016/j.ribaf.2021.101554.

Game-theoretic models in multi-agent systems: emergent behaviours, critical phase transactions and learning dynamics

Supervisor: Dr. Stefanos Leonardos

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, Game Theory, Foundations of computing (algorithms, computational complexity)

Project Description

This project is aimed at students who are interested in cutting-edge research at the intersection of multi-agent systems, game theory and learning dynamics, with applications in economics, machine learning, and artificial intelligence. The project's objective is to explore the intricate patterns of multi-agent systems through a game-theoretic lens, emphasizing on learning dynamics, chaos theory, and their applications. Special focus will be placed on understanding the emergent behaviors in algorithmic decision-making processes that continuously evolve over time. The study will explore phase transitions in strategic interactions, analyze or develop novel algorithms, and quantify their implications on coordination and competition in real-world systems. The analysis will use tools from game theory, mathematics and the theory of dynamical systems, to develop, study and apply learning algorithms in complex multi-agent systems. Successful applicants will have the chance to shape the future of learning systems, bridging theoretical advancements with practical applications with the frameworks of machine learning and artificial intelligence.

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.

Sustainable Delivery Logistics with Drones and Cargo Bikes

Supervisor: Dimitrios Letsios

Areas: Robotics, Foundations of computing (algorithms, computational complexity), Artificial Intelligence (symbolic AI, logic, etc.)

Project Description

Effectively solving inventory optimisation, scheduling, and vehicle routing problems is essential to meet customer demands and manage fleet, human worker, and storage costs for companies in the logistics sector, e.g. courier and delivery services. Drones and cargo bikes offer an alternative to cost-effective and sustainable last-mile deliveries in urban environments. The goal of this project is to develop data science and optimisation methodologies (e.g. metaheuristics and integer programming) for generating solutions to challenging logistics optimisation problems with drones and cargo bikes that attain optimal trade-offs between solution quality and computational effort, while enhancing their performance with simulations and machine/deep learning. Proposed approaches will be evaluated using recent data and benchmark case studies from business partners and in the literature. The project will include a thorough study on effective hyper-parameter tuning (e.g. initialisation components, solution space structure, search operators) of optimisation methodologies for computing the best possible solution within given time frames. Further, it will investigate innovative ways of enhancing performance with dynamic parameter adaptations and machine/deep learning, based on information (e.g. rewards, data properties) collected during the solution process. The project will also assess and strengthen the robustness of proposed approaches via simulations.

Goal-based explanations for autonomous systems and robots

Supervisor: Gerard Canal

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Robotics

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 that 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 ongoing task. In this case, the new plan may be different to the original one, thus changing the behaviour that the robot was performing. This project will explore how to generate goalbased 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.

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).

Leveraging Language Models for Contextual Vulnerability Identification

Supervisor: Maher Salem

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, Cybersecurity, Systems (software engineering, programming)

Project Description

As software systems grow increasingly complex, the need for effective vulnerability detection methods becomes paramount. Traditional static analysis tools often struggle to identify context-specific vulnerabilities due to their reliance on predefined patterns and rules. This research proposes leveraging advanced language models, such as transformers, to enhance the identification of vulnerabilities in software code by understanding its context. The central idea of this topic is to explore how large language models (LLMs) can be trained to analyze code not merely as isolated snippets but as part of a larger context. By fine-tuning LLMs on extensive datasets that include both vulnerable and secure code, the model can learn to recognize subtle patterns and interactions that indicate potential vulnerabilities. This approach aims to move beyond conventional methods by incorporating an understanding of how different code components interact with each other, thereby improving detection accuracy. The research will involve several key phases. First, a comprehensive dataset will be curated, containing various programming languages and a range of vulnerability types, such as SQL injection, cross-site scripting, and buffer overflows. This dataset will serve as the foundation for training the language models. Next, the study will focus on developing a framework that integrates the LLMs into an existing vulnerability detection pipeline, allowing for real-time analysis and feedback during the software development lifecycle. Furthermore, the research will explore the effectiveness of different model architectures and training techniques, including transfer learning and few-shot learning, to optimize performance. By evaluating the models against established benchmarks and real-world codebases, the study aims to quantify improvements in vulnerability detection rates compared to traditional static analysis tools. Another important aspect of this research is the interpretability of the model's predictions. It is crucial for developers to understand why a particular piece of code was flagged as potentially vulnerable. Therefore, the study will investigate methods to enhance the transparency of LLMs, providing explanations that can guide developers in addressing identified vulnerabilities. Ultimately, this research seeks to contribute to the field of cybersecurity by providing a novel approach to vulnerability detection that leverages the capabilities of modern AI. By harnessing the contextual understanding of language models, the goal is to create more robust and intelligent tools that can significantly enhance software security, helping developers proactively identify and mitigate vulnerabilities before they can be exploited.

References

DOI: 10.1145/3460318.3464820
DOI: 10.1145/3404835.3462831
DOI: 10.1145/3397670
DOI: 10.1109/TSE.2020.2986860

Advanced Modelling on Multimodal Urban Geospatial Data Fusion - Case Studies for UK Cities

Supervisor: Yijing Li

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, urban data science

Project Description

The project aims to set up a framework applying interdisciplinary advanced model(s) onto multimodal data, especially urban geospatial datasets collected in UK cities, to realise least-uncertainty data fusion and integration. Throughout the project, external partners proposed practical case study projects will be utilised to test the model(s) performance with expectation of wider research impacts into real urban applications. Archived datasets compiled at CUSP London will be provided for project kicking off, including transport/mobility, crime, environment, health, air quality, economy and demographics statistics; the candidate is expected to apply data mining techniques to collect other multimodal urban datasets such as scene images, social media records, etc., be resilient to learn and apply multi-disciplinary methods, and be confident to translate research outputs into policy-inform languages.

Exploring Interactive Multi-Dimensional Approaches of Delivery of Communication in Patient Scenarios in Oral Health Education

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

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Natural Language Processing, Human-centred computing (human-computer interaction), Machine learning / Deep learning, Education

Project Description

This is a joint project between the Department of Informatics and the Faculty of Dentistry, Oral & Craniofacial Sciences (FoDOCS). Communication in patient scenarios in oral health education can be cost-intensive in terms of time and resources. In this project, we propose exploring interactive multidimensional approaches such as immersive technology, text-to-text, and voice-to-voice communication delivery in patient scenarios in oral health education. These approaches enhance the learning experience and offer a cost-effective solution, making the delivery of communications in patient scenarios in oral health education more feasible and sustainable. This project aims to design and create adaptable, contextually relevant patient scenarios, offering engaging and realistic interactions for students. The beauty of these approaches is their adaptability. Whether it is a VR interactive tool, text-based, or voice-based conversation, they can all respond to students' inquiries and actions in real-time, mimicking the interaction they would normally have in the clinic. This adaptability ensures the relevance and effectiveness of the project in various educational settings. We aim to examine the Generative Language Models (GLMs) to generate customized case studies and simulation scenarios so that each learner can practice specific skills repeatedly in a controlled environment. This encourages the acquisition and refinement of skills, such as explaining the importance of oral hygiene and discussing dietary habits. The critical focus is patient-clinician communication, behaviour change, professionalism, etc.

Software sustainability analysis and improvement

Supervisor: Kevin Lano

Areas: Systems (software engineering, programming), Machine learning / Deep learning, Artificial Intelligence (symbolic AI, logic, etc.)

Project Description

The project would consider techniques for analysing software sustainability (in the sense of energy use and energy efficiency) using rule-based analysis and refactoring, or by the use of deep learning techniques such as LLMs to identify energy use flaws and potential refactorings. It would be particularly useful to consider analysis and refactorings at the specification or design levels of a software system, in order that programming-language independent advice and improvements can be made. 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.

Safe Reinforcement Learning from Human Feedback

Supervisor: Yali Du

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning

Project Description

Reinforcement learning (RL) has become a new paradigm for solving complex decision making problems. However, it presents numerous safety concerns in real world decision making, such as unsafe exploration, unrealistic reward function, etc. As reinforcement learning agents are frequently evaluated in terms of rewards, it is less noticed that designing AI agents that have the capability to achieve arbitrary objectives can be deficient in that the systems are intrinsically unpredictable and might result in negative and irreversible outcomes to humans. While humans understand the dangers, human involvement in the agent's learning process can be promising to boost AI safety for being more aligned with human values [1]. Dr. Du's early research [2,3] shows that human preference can be used as an effective replacement for reward signals. One recent attempt [1] also adopted human preference as a replacement for reward signals, to guide the training of agents in safety-critical environments; while agents query humans with a certain probability, how to actively query humans and adapt its knowledge to the task and query is not considered. This project considers to build safe RL agents leveraging human feedback, and aims to address two challenges: 1) how to enable agents to actively query humans with efficiency thus minimising disturbance to humans; 2) how to improve algorithms' robustness in dealing with large state space and even unseen tasks. The target of this project is to realise human value alignment safe RL in a scalable (in terms of task scale) and efficient (in terms of human involvement) way.

References

[1] A review of safe reinforcement learning: Methods, theory and applications. Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. IEEE TPAMI 2024.

[2] Meta-Reward-Net: Implicitly Differentiable Reward Learning for Preference-based Reinforcement Learning. Runze Liu, Fengshuo Bai, Yali Du, Yaodong Yang. NeurIPS 2022

[3] Safe Reinforcement Learning with Free-form Natural Language Constraints and Pre-Trained Language Models. Xingzhou Lou, Junge Zhang, Ziyan Wang, Kaiqi Huang, Yali Du. AAMAS 2024

Multi-agent Cooperation with RL and LLMs

Supervisor: Yali Du

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, Natural Language Processing, Robotics

Project Description

Multi-agent systems (MAS) have become increasingly relevant in fields such as robotics, finance, and autonomous systems. However, achieving effective cooperation among multiple agents remains challenging, especially in dynamic and uncertain environments. RL has been a powerful method for training agents, but traditional approaches often struggle with scalability and communication bottlenecks. Meanwhile, LLMs have demonstrated remarkable capabilities in language understanding and generation, which can be leveraged to facilitate communication and strategy development among agents. This study aims to explore how reinforcement learning (RL) can be combined with large language models (LLMs) to improve multi-agent cooperation in complex environments. The goal is to enhance communication, decision-making, and coordination between agents, enabling them to solve tasks that require a high level of collaboration and safety. This project explores the questions of 1) How can LLMs be integrated into multi-agent systems to enhance cooperation and communication among agents trained using RL? 2) What are the optimal communication protocols that maximize the synergy between LLMs and RL in multi-agent scenarios? 3) How can this combination be scaled to large numbers of agents while maintaining efficiency and performance? Dr Du's early attempts explored how to leverage LLMs for communication, and incorporated human instructions to ensure safe and cooperative control, with examples including the game of Werewolf, football, and safe robot control. This research will contribute to the field of multi-agent systems by developing new techniques for improved cooperation using cutting-edge LLMs. The findings could be applicable in various industries, including autonomous vehicles, robotics, and distributed AI systems, where multi-agent cooperation is critical for success.

References

[1] Safe Multi-agent Reinforcement Learning with Natural Language Constraints. Ziyan Wang, Meng Fang, Tristan Tomilin, Fei Fang, Yali Du. Arxiv 2024.

[2] Understanding, Rehearsing, and Introspecting: Learn a Policy from Textual Tutorial Books in Football Games. Xiong-Hui Chen, Ziyan Wang, Yali Du, Meng Fang, Shengyi Jiang, Yang Yu, Jun Wang. NeurIPS 2024 Oral.

[3] Learning to Discuss Strategically: A Case Study on One Night Ultimate Werewolf. Xuanfa Jin, Ziyan Wang, Yali Du, Meng Fang, Haifeng Zhang, Jun Wang. NeurIPS 2024.

Assessing the value of evidence with argument-driven credal networks

Supervisor: Jeroen Keppens

Areas: Artificial Intelligence (symbolic AI, logic, etc.)

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.

Investigating LLM-based Generative AI Applications in Cybersecurity

Supervisor: Ievgeniia Kuzminykh/Hannan Xiao

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Cybersecurity, Natural Language Processing, Human-centred computing (human-computer interaction)

Project Description

The rapid development and deployment of GPT-based agents in cybersecurity mark a significant leap forward in approaching digital security challenges from a practitioner standpoint. Under this topic you can explore the ways generative AI is impacting the cybersecurity industry, from both sides, such as gen AI for security and security of gen AI. From one side, malicious attackers are seizing the potential of generative AI to launch cyber attacks that are harder to detect and defend against. OWASP top 10 for Generative AI [1] lists out the top 10 vulnerabilities impacting the applications usilising LLM. Prompt Injection, Insecure Output Handling and Data Poisoning take the top 3 spots and are also the root causes for the other type of vulnerabilities (Overreliance, Insecure Plugins etc) as shown in research paper [2]. From another side, Gen AI is also helping make security teams more accurate, efficient, and productive in defending their organisations. Examples of utilising of generative AI for security operations could be [3] : - Supplementing understaffed security teams - Detecting threats in real time - Improving incident response. The potential topics in this project area could include but not limited to: 1. Optimisation of prompts for security related topics. Through clever prompt engineering (called jailbreaking [4]), LLMs can be made to reveal internal mechanisms, share private data, produce offensive speech, or perform unintended workloads. LLMs thus pose a security risks [1, 5, 6]. 2. Prompt injection detection mechanisms. 3. Ensuring online safety using LLM. AI seems like the perfect response to the growing challenges of content moderation on social media platforms: the immense scale of the data, the relentlessness of the violations, and the need for human judgments without wanting humans to have to make them. The paper [7] elaborates on the topic of prompt/response classifiers. The prompts and answers could be classified into groups such as safe and harmful. Typical examples of a harm would be Child Safety, Exfiltrating PII/SPII, Sexually Explicit Content, Malicious/Dangerous content. 4. Content moderation using LLM. Similar to previous but can be extended to the detection of harassment and throlling [8]. 5. Understanding Generative AI for Cloud Security. Generative AI can make new data from existing patterns. For cloud security, this means it has the potential to: 6a. Simulate Threat Scenarios: Generative AI can create realistic threat scenarios, allowing security teams to test and validate their Cloud infrastructure's resilience. By simulating potential attack vectors, organizations can proactively identify vulnerabilities and take steps to ensure they are protected against them before they are exploited. 6b. Optimize Security Configurations: AWS offers a number of services, each with its own set of security configurations. With Generative AI, we can analyze existing configurations, simulate various combinations, and ask Generative AI to provide recommendations based on our specific needs. 6c. Enhance Monitoring and Alerts: By training on historical security logs and events, Generative AI can predict potential security breaches or anomalies. The key word here is "potential." Knowing what "could" happen allows security teams time to prepare and allows for more rapid action to be taken. 6. Understanding Generative AI for firewall optimisation Generative AI could simulate web traffic patterns based on your historical log data and compare that to your existing WAF or firewall rules, ensuring that malicious requests are blocked while legitimate traffic flows seamlessly. 7. 10. Understanding of Gen AI for Qualitative audit of security policies Each organisation is governed by a security policy, which technically or conceptually specifies a number of guidelines for ensuring IT security. You will investigate whether GenAI can be employed to translate a security policy for wider staff [9, 10]..

References

[1] https://owasp.org/www-project-top-10-for-large-language-model-applications/
[2] Knowledge Bases and Language Models: Complementing Forces. https://link.springer.com/chapter/10.1007/978-3-031-45072-3_1
[3] https://secureframe.com/blog/generative-ai-cybersecurity
[4] Liu, Y., et al.: Jailbreaking ChatGPT via prompt engineering: an empirical study. arXiv preprint arXiv:2305.13860 (2023)
[5] https://simonwillison.net/2023/May/2/prompt-injection-explained/
[6] https://www.geeksforgeeks.org/what-is-jailbreak-chat/
[7] Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations https://arxiv.org/html/2312.06674v1
[8] Content moderation, AI, and the question of scale https://journals.sagepub.com/doi/full/10.1177/2053951720943234
[9] Introducing Rules Genie: Generative AI for Automating Policy Creationhttps://www.opsmx.com/blog/introducing-rules-genie-generative-ai-for-automating-policy-creation/
[10] The LLM Police: Between Firewalls and Policies for Generative AI https://www.fairly.ai/blog/the-llm-police-between-firewalls-and-policies-for-generative-ai

Argument mining

Supervisor: Oana Cocarascu

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, Natural Language Processing

Project Description

In everyday life, decisions are often based on arguments, counter-arguments, and facts. While arguments are claims backed by reasons that are supported by evidence, facts can be proven with clear and objective data. Automatically identifying and presenting facts and arguments can not only facilitate and challenge debates, but also aid humans and automated systems in reaching decisions, hence the societal impact of this task is tremendous.

Computational argumentation is a research area in natural language processing which encompasses several tasks such as argument mining, argument reasoning, and argument generation amongst others. Much progress has been made in recent years on argument mining whereby the task is to determine whether a text represents an argument, followed by identifying the arguments for or against an issue. Argument mining has been applied to several areas: persuasive essays, scientific articles, Wikipedia articles, news articles, online debates, product reviews, social media, legal documents, and political debates.

The project aims to develop computational methods that find, extract, and evaluate arguments in text as well as deal with incomplete arguments, i.e. arguments that can be understood using background knowledge.

References

[1] https://aclanthology.org/2022.tacl-1.80.pdf
[2] https://aclanthology.org/2024.acl-long.126.pdf

Multilingual argument mining

Supervisor: Oana Cocarascu

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning, Natural Language Processing

Project Description

In everyday life, decisions are often based on arguments, counter-arguments, and facts. While arguments are claims backed by reasons that are supported by evidence, facts can be proven with clear and objective data. Automatically identifying and presenting facts and arguments can not only facilitate and challenge debates, but also aid humans and automated systems in reaching decisions, hence the societal impact of this task is tremendous.

Computational argumentation is a research area in natural language processing which encompasses several tasks such as argument mining, argument reasoning, and argument generation amongst others. Much progress has been made in recent years on argument mining whereby the task is to determine whether a text represents an argument, followed by identifying the arguments for or against an issue. Argument mining has been applied to several areas: persuasive essays, scientific articles, Wikipedia articles, news articles, online debates, product reviews, social media, legal documents, and political debates.

Despite the growing interest in computational argumentation, the majority of datasets are in English. The project will focus on argument mining in low-resource languages and will develop novel corpora and algorithms for multilingual argument mining.

References

[1] https://aclanthology.org/2022.tacl-1.80.pdf
[2] https://aclanthology.org/2024.acl-long.126.pdf
[3] https://aclanthology.org/2024.acl-long.628.pdf
[4] https://aclanthology.org/2020.findings-emnlp.29.pdf

AI in finance

Supervisor: Carmine Ventre

Areas: Artificial Intelligence (symbolic AI, logic, etc.)

Project Description

AI in finance, AI tools have been developed, and often successfully deployed, for many tasks in different application domains. The stochastic nature of finance, however, makes it hard to safely adopt the ideas underpinning much of the progress in the area. It is not even clear when AI is safe to adopt in this domain. Whilst high profits for a variety of backtesting scenarios can be a proxy in the case of trading AI bots, for other tasks, such as, pricing derivatives or defining risk models, the connoisseur approach tends to be favoured. Finally, there are areas within finance which beg for the use of AI to add safety and trustworthiness, the primary example being Environmental, social, and corporate governance (ESG) where often data is easily manipulated (a.k.a., greenwashing). In this project, we will focus on the design of AI to address these issues. The translation of AI techniques to the uncertain and often adversarial financial applications will consider safety and trustworthiness as a first-class concern. Exemplar research directions include the following. Mathematical approaches, including the study of the latent space of AI models, can provide explanaibility and substitute experts with formal guarantees. The computational efficiency of symbolic approaches, like Monte Carlo simulations, needs to be compared with more modern model-based deep learning methods to add speed as a measure of trustworthiness to the high-paced world of finance. The adoption of symbolic techniques, such as intrinsic-time framework, should be adopted for the labelling of financial data, as opposed to the approaches which simply trust the noisy market data. Transformers for NLP can be used to leverage textual resources and assess the ESG rating of financial products and/or firms. Prospective applicants are encouraged to consult the publications of Prof Ventre at https://kclpure.kcl.ac.uk/portal/en/persons/carmine.ventre/publications/.

Adaptation and effective communication in collaborative physically Assistive Tasks

Supervisor: Gerard Canal

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Robotics

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.

Explaining robotic planning decision points along execution

Supervisor: Gerard Canal

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Robotics

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 on current 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 it 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.

Enhancing Safety in Robotics by Tackling Blind-Spots and Bias in AI Models

Supervisor: Gerard Canal

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Robotics

Project Description

The current revolution of artificial intelligence (AI) is becoming more prominent and its potential is still to be unleashed. In the context of robotics, AI can provide support to multiple scenarios, among them, industry, education and healthcare. It is important to know how these systems can work on these contexts but it is imperative that they can treat people respectfully and equally. There are significant efforts in this direction that focus on the context of fairness and explainability. Several AI models normally employed in robotics, such as computer vision models, have been tested to discover that they still contain blind-spots in their detection capabilities, several of them affecting specifically protected groups, such as children or citizens with disabilities. Even if the models are becoming more explainable these days, the consequences of these blind-spots in their explanations and especially the actions of the robots in the real world still requires deeper studies. This is particularly important due to the safety issues that this may impose, which is specially critical in assistive scenarios where a robot helps a user from a vulnerable group perform activities of daily living. This thesis aims to address these issues by: 1) Identifying use cases where the sensitiveness of fairness issues might have a strong repercussion in the behaviour of the robots, with a special emphasis on when this results in unsafe situations for the user recipient of the assistance. This will consist of collecting different examples for the literature that the student can have access and implementing them with the robots that we have available in the department such as the PAL Robotics' TIAGo or models of smart cars. It will also potentially employ digital twins to create a simulation environment for more complex robots. 2) Create strategies to identify blind-spots. Based on the previous work of adversarial machine learning where blind-spots are normally identified as misclassifications or mis-actions that a robot will execute, this part of the thesis will work on identifying and designing adversarial scenarios that will make the system misbehave. The scenario design will consider potential sensory alterations that the robot will face, especially connected with environment conditions. With this information, the thesis will aim to explain the scenario and the specific conditions that led to the misclassification. This will support redesigning the learning process and will serve for standardising benchmark testing conditions. 3) Based on the previous adversarial scenarios and the specific transformations that led the system to make erroneous decisions, this last part will provide explanations about the system limitations, with an aim to enhance the safety of the system. It will focus on: 1) generalising from the adversarial scenarios to create explanations and 2) inverse the pipeline and create adversarial conditions from specific explanations. These adversarial conditions will be focused on fairness. Besides this last part will put a strong effort on evaluating explanatory systems for robotics under adversarial conditions.

References

Canal, G., Torras, C., & Alenya, G. (2021). Are preferences useful for better assistance? a physically assistive robotics user study. ACM Transactions on Human-Robot Interaction (THRI), 10(4), 1-19.

Vila Abad, M., Canal, G., & Alenya, G. (2018). Towards safety in physically assistive robots: eating assistance. In: Proceedings of the 2018 IROS Workshop on Robots for Assisted Living

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.

Menendez, H. D., Bhattacharya, S., Clark, D., & Barr, E. T. (2019). The arms race: Adversarial search defeats entropy used to detect malware. Expert Systems with Applications, 118, 246-260.

Calleja, A., Martin, A., Menendez, H. D., Tapiador, J., & Clark, D. (2018). Picking on the family: Disrupting android malware triage by forcing misclassification. Expert Systems with Applications, 95, 113-126.

Menendez, H. D. (2022, October). Measuring Machine Learning Robustness in front of Static and Dynamic Adversaries. In 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 174-181). IEEE.

Agents powered by foundation models

Supervisor: Helen Yannakoudakis

Areas: Machine learning / Deep learning, Artificial Intelligence (symbolic AI, logic, etc.), Natural Language Processing

Project Description

With the expansive capabilities of foundation models, the concept of building agents powered by these models (like large language models) has recently emerged. Several demonstration projects, such as AutoGPT, GPT-Engineer, and BabyAGI, illustrate this potential. Foundation models offer possibilities beyond creating images, well-crafted text, stories, essays, and code—they can serve as powerful general problem solvers. This project aims to develop agents driven by foundation models that can observe, take action, and respond to feedback in a continuous loop with external environments, including interactions with humans, tools, and the physical world. The focus will be on two key areas: specialization and multi-modality.

Causal Explanations for Sequential Decision Making

Supervisor: Nicola Paoletti

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning

Project Description

Explainable AI has become increasingly relevant, because in many domains, especially safety-critical ones, it is desirable to complement black-box machine learning (ML) models with comprehensible explanations of the models' predictions. This project focuses on explanations for sequential decision making processes. Such processes are found in AI planning, reinforcement learning, and control/cyber-physical systems, and they nowadays make use of ML models to e.g., represent the policy or the environment's dynamics. Unlike most explainability techniques that deal with input-output, i.e., one-step, predictions, the challenge here is to deal with sequence data that arise from multiple, inter-dependent steps taken over time. Moreover, explanations need to account for the uncertain or probabilistic environment dynamics. In particular, the focus will be on causal explanations building on the actual causality framework by Halpern and Pearl [1,2]. Given a realization of the sequential process under study, we seek to find the minimal set of units (e.g., observed steps, policy actions, agents) responsible for the observed outcome, i.e., such that the counterfactual model obtained by changing such units leads to a different outcome. We welcome project proposals around any of the following topics (or similar) that our group is currently investigating:

- Counterfactual Inference of Markov Decision Processes [3-6]
- Dealing with uncertain models, partial observability, unobserved confounders [7,8]
- Combining counterfactuals with temporal logic reasoning for verification [9-11]
- Reliable counterfactual inference with data-driven models [12,13]

References

[1] Halpern, Joseph Y., and Judea Pearl. "Causes and explanations: A structural-model approach. Part II: Explanations." The British journal for the philosophy of science (2005).
[2] Beckers, Sander. "Causal explanations and XAI." Conference on Causal Learning and Reasoning. PMLR, 2022.
[3] Oberst, Michael, and David Sontag. "Counterfactual off-policy evaluation with gumbel-max structural causal models." International Conference on Machine Learning. PMLR, 2019.
[4] Tsirtsis, Stratis, Abir De, and Manuel Rodriguez. "Counterfactual explanations in sequential decision making under uncertainty." Advances in Neural Information Processing Systems 34 (2021): 30127-30139.
[5] Lorberbom, Guy, et al. "Learning generalized gumbel-max causal mechanisms." Advances in Neural Information Processing Systems 34 (2021): 26792-26803.
[6] Triantafyllou, Stelios, Adish Singla, and Goran Radanovic. "Actual causality and responsibility attribution in decentralized partially observable Markov decision processes." Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 2022.
[7] Lu, Chaochao, Bernhard Scholkopf, and Jose Miguel Hernandez-Lobato. "Deconfounding reinforcement learning in observational settings." arXiv preprint arXiv:1812.10576 (2018).
[8] Zhang, Junzhe, and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical Report R-23, Purdue AI Lab, 2016.
[9] Kazemi, Milad, and Nicola Paoletti. "Causal Temporal Reasoning for Markov Decision Processes." arXiv preprint arXiv:2212.08712v2 (2023).
[10] Finkbeiner, Bernd, and Julian Siber. "Counterfactuals modulo temporal logics." arXiv preprint arXiv:2306.08916 (2023).
[11] Coenen, Norine, et al. "Temporal causality in reactive systems." International Symposium on Automated Technology for Verification and Analysis. Cham: Springer International Publishing, 2022.
[12] Chernozhukov, V., Wuthrich, K., & Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. Journal of the American Statistical Association, 116(536), 1849-1864.
[13] Lei, L., & Candes, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83(5), 911-938

Reliable Learning for Safe Autonomy with Conformal Prediction

Supervisor: Nicola Paoletti

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning

Project Description

For their high expressive power and accuracy, machine learning (ML) models are now found in countless application domains. These include autonomous and cyber-physical systems found in high-risk and safety-critical domains, such as healthcare and automotive. These systems nowadays integrate multiple ML components for e.g., sensing, end-to-end control, predictive monitoring, anomaly detection. Hence, data-driven analysis has become necessary in this context, one where rigourous modeldriven techniques like model checking have been the go-to solution for years. In this project you will develop data-driven analysis techniques for autonomous systems based on conformal prediction (CP) [1,2], an increasingly popular approach to provide guarantees on the generalization error of ML models: it can be applied on top of any supervised learning model and it provides so-called prediction regions (instead of single-point predictions) guaranteed to contain the (unknown) ground truth with given probability. Crucially, these coverage guarantees are finite-sample (as opposed to asymptotic) and do not rely on any parametric or distributional assumptions. Our group has a track record of developing CP-based methods for predictive monitoring of autonomous and cyber-physical systems [3-6]. With this project, you will contribute to this endeavour working on challenge problems including off-policy prediction [7,8], data-driven optimization, causal inference [9,10], robust inference under distribution shifts [11,12] and uncertain distributions [13,14]. The proposed techniques will be evaluated in standard relevant benchmarks and different real-world scenarios coming from the REXASI-PRO EU project [15], which focuses on safe navigation of autonomous wheelchairs in crowded environments for people with reduced mobility.

References

[1] Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Vol. 29. New York: Springer, 2005.
[2] Angelopoulos, Anastasios N., and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv preprint arXiv:2107.07511 (2021).
[3] Cairoli, Francesca, Nicola Paoletti, and Luca Bortolussi. "Conformal quantitative predictive monitoring of STL requirements for stochastic processes." Proceedings of the 26th ACM International Conference on Hybrid Systems: Computation and Control. 2023.
[4] Cairoli, Francesca, Luca Bortolussi, and Nicola Paoletti. "Learning-Based Approaches to Predictive Monitoring with Conformal Statistical Guarantees." International Conference on Runtime Verification. Cham: Springer Nature Switzerland, 2023.
[5] Bortolussi, Luca, et al. "Neural predictive monitoring and a comparison of frequentist and Bayesian approaches." International Journal on Software Tools for Technology Transfer 23.4 (2021): 615-640.
[6] Cairoli, Francesca, Luca Bortolussi, and Nicola Paoletti. "Neural predictive monitoring under partial observability." Runtime Verification: 21st International Conference, RV 2021, Virtual Event, October 11—14, 2021, Proceedings 21. Springer International Publishing, 2021.
[7] Russo, Alessio, Daniele Foffano, and Alexandre Proutiere. "Conformal Off-Policy Evaluation in Markov Decision Processes." 62nd IEEE Conference on Decision and Control, Dec. 13-15, 2023, Singapore. IEEE, 2023.
[8] Taufiq, Muhammad Faaiz, et al. "Conformal off-policy prediction in contextual bandits." Advances in Neural Information Processing Systems 35 (2022): 31512-31524.
[9] Lei, L., & Candes, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83(5), 911-938.
[10] Chernozhukov, V., Wuthrich, K., & Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. Journal of the American Statistical Association, 116(536), 1849-1864.
[11] Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2), 816-845.
[12] Gibbs, Isaac, and Emmanuel Candes. "Adaptive conformal inference under distribution shift." Advances in Neural Information Processing Systems 34 (2021): 1660-1672.
[13] Cauchois, M., Gupta, S., Ali, A., & Duchi, J. C. (2020). Robust validation: Confident predictions even when distributions shift. arXiv preprint arXiv:2008.04267.
[14] Gendler, A., Weng, T. W., Daniel, L., & Romano, Y. (2021, October). Adversarially robust conformal prediction. In International Conference on Learning Representations.
[15] REliable & eXplAinable Swarm Intelligence for People with Reduced mObility (REXASI-PRO), https://rexasipro.spindoxlabs.com/.

Planning and Reinforcement Learning for versatile autonomous robots

Supervisor: Matteo Leonetti

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Robotics

Project Description

Model-based reinforcement learning has been lagging behind initial and exciting model-free results in deep reinforcement learning. In this project we will consider the problem of an autonomous robot required to carry out different tasks in its environment, frequently switching between goals. The research will focus on model learning and effective use of models to drive exploration, hierarchical models, and multi-task heuristics. Examples of previous work in this direction are provided in the reference section

References

1. Probabilistic Inference of Human Capabilities from Passive Observations. Peter Tisnikar, Gerard Canal, Matteo Leonetti. Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024.
2. Online human capability estimation through reinforcement learning and interaction. Chengke Sun, Anthony G Cohn, Matteo Leonetti. Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023.
3. Reducing the Planning Horizon Through Reinforcement Learning. Logan Dunbar, Benjamin Rosman, Anthony G Cohn, Matteo Leonetti. Proc. of Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), 2022.
4. Meta-Reinforcement Learning for Heuristic Planning. Ricardo Luna Gutierrez, Matteo Leonetti. Proc. of International Conference on Automated Planning and Scheduling (ICAPS), 2021.
5. A synthesis of automated planning and reinforcement learning for efficient, robust decision-making. Matteo Leonetti, Luca Iocchi, Peter Stone. Artificial Intelligence, 2016.

Designing and Developing a framework for responsible security and privacy practices for GenAI Tools

Supervisor: Maher Salem

Areas: Cybersecurity, Human-centred computing (human-computer interaction), Artificial Intelligence (symbolic AI, logic, etc.)

Project Description

Generative Artificial Intelligence (GenAI) technologies have transformed human life, impacting areas such as healthcare, education, and social interactions. While GenAI tools offer creative content generation, data synthesis, and automation benefits, they also introduce significant challenges that may negatively impact individuals and communities, particularly vulnerable groups. These challenges include privacy and ethical issues, legal risks, bias and discrimination, misinformation, and inaccurate outputs. Despite the importance of these issues, there is limited empirical research on users' experiences and views regarding GenAI security and privacy. This project aims to apply both qualitative and quantitative methods to investigate how users interact with GenAI tools, the reasons behind their use, and how these experiences shape perceptions of GenAI. A key component of the study will involve understanding users' mental models of the benefits and risks of GenAI, the educational resources they use to assess potential risks, and the protective measures they adopt. By examining users' learning processes and resources, this research will provide insights into gaps in GenAI literacy. The study's educational goals include developing targeted resources and practical guidelines to improve GenAI literacy among diverse groups. Empirical insights from this research will guide the design of safeguards for GenAI technologies and inform curriculum and policy recommendations, enabling institutions to equip students, educators, and users with the skills to navigate GenAI responsibly and securely.

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

Supervisor: Frederik Mallmann-Trenn

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning

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

Discovering the Secrets of Random Neural Networks - Training by Pruning

Supervisor: Frederik Mallmann-Trenn

Areas: Artificial Intelligence (symbolic AI, logic, etc.), Machine learning / Deep learning

Project Description

Deep learning has revolutionized many fields like, most prominantely, natural language processing, where large language models such as ChatGPT and Gemini represent a groundbreaking advancement. However, these progresses come at a significant energy cost due to the massive number of connections (trillions) in neural networks (NNs). Although the cost of each connections is small, the sheer number of them results in enormous costs: The inference cost of each query to ChatGPT-4 is estimated to cost $0.34. They key to reducing the inference cost is thus to to reduce the number of parameters (connec- tions). We aim to do precisely that. More precisely, the goal of this fellowship is to obtain algorithms for the sparsification of neural networks - reducing the number of parameters by orders of magnitude. The goal of this project is to attain energy savings by relying on training by pruning. In the simplest case [2], we are given a target network N and we initialise a network N' with random weights. The goal then is to remove edges from N' as to approximate N. Recently (e.g. [2]) as shown that this is always possible provided that N' is sufficiently large. They don't show however how the network N' can be found - they only prove the existence. The goal of the PhD will be to find such networks efficiently. Some methods have been proposed (e.g., [1]), but so far no proof is known, which this projects aims to change. The impact of this could be huge. It is absolutely necessary to have a strong math and stats background.

References

[1] Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A. and Rastegari, M., 2020. What's hidden in a randomly weighted neural network?. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11893-11902). https://openaccess.thecvf.com/content_CVPR_2020/papers/Ramanujan_Whats_Hidden_in_a_Randomly_Weighted_Neural_Network_CVPR_2020_paper.pdf
[2] Malach, E., Yehudai, G., Shalev-Schwartz, S. and Shamir, O., 2020, November. Proving the lottery ticket hypothesis: Pruning is all you need. In International Conference on Machine Learning (pp. 6682-6691). PMLR. https://proceedings.mlr.press/v119/malach20a/malach20a.pdf