Informatics PhD projects at King's College London, AY 24-25 — Artificial Intelligence

The PhD project proposals listed below will be considered for 2024/25 studentships available in the Department of Informatics to start 1 October 2024 or later during the 2024/25 academic year. Please note that this list is not inclusive and potential applicants can alternatively identify and contact appropriate supervisors outlining their background and research interests or proposing their own project ideas.
The PhD projects are listed in two groups. In the first group are the projects with allocated studentships: each project in this group has one allocated studentship. The remaining studentships will be considered for the projects listed in the second group. The number of those remaining studentships is smaller than the number of the projects in the second group. The allocation of studentships will be based on the merits of individual applications. Applications for PhD studies in the Department of Informatics, for all listed projects as well as for other projects agreed with supervisors, are also welcome from students applying for other funding (within other studentship schemes) and from self-funded students. See also this list of funding opportunities available at King's for post-graduate research in Computer Science.

- Scholarship Allocated
- Scholarship Not Allocated

Scholarship Allocated

(Back to Top)

- Goal-based explanations for autonomous systems and robots
- Common Sense Planning (for Robotics)
- Adaptation and effective communication in collaborative physically Assistive Tasks
- Explaining robotic planning decision points along execution
- Reliability and verification of software for scientific and ML computing
- Towards Robust Reasoning in Large Language Models
- Robotics and Social Justice
- Synthetic video generation: counterfactual explanations
- Improving Understandability of Automatically Generated Test Cases using Text-to-Text Transformer Models
- Planning and Reinforcement Learning for versatile autonomous robots
- Safe Reinforcement Learning from Human Feedback
- Unifying Principals in Safe and Trusted Assistive AI
- Safe Reinforcement Learning from Human Feedback
- Estimating the ground truth of LLMs in softare engineering Tasks
- Text promptable surgical video generation

Goal-based explanations for autonomous systems and robots

Supervisor: Gerard Canal

Areas: Artificial Intelligence (AI), Human-centred computing

(Back to Scholarship Allocated)

Project Description

Autonomous systems such as robots may become another appliance found in our homes and workplaces. In order to have such systems helping humans to perform their tasks, they must be as autonomous as possible, to prevent becoming a nuisance instead of an aid. Autonomy will require the systems or robots to set up their own agenda (in line with the tasks they are meant to do), defining the next goals to achieve and discarding those 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 goal-based explanations for robots in assistive/home-based scenarios, extracted from goal-reasoning techniques. It will also look at plan repair to enforce cohesion after a replanning to ideally increase the trust and understanding of the users about the system. Those explanations should also contemplate unforeseen circumstances, therefore explaining things based on "excuses" that the robot may give to the user. Finally, we will investigate how to obtain and provide those explanations at execution time, so explaining on the go. The methods developed shall be integrated into a robotic system, in an assistive/service robot scenario.

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).

Common Sense Planning (for Robotics)

Supervisor: Gerard Canal / Albert Merono-Penuela

Areas: Artificial Intelligence (AI), Human-centred computing

(Back to Scholarship Allocated)

Project Description

Task Planning (also known as Symbolic Planning or AI Planning) has proved to be a very useful technique to tackle the decision-making problem in robotics. Given a set of task goals, the planner can come up with a set of actions that will reach those goals once executed by the robot. However, plans are often short-lived when robots execute them, given that the real world is complex, and some actions may fail. A traditional approach is to recompute plans once they fail (replanning), however, computing new plans is often costly. This is an issue in robotics and real-time systems, where users wouldn't want a robot that stops for some minutes to compute a plan every now and then. Instead of replanning, a solution could be to repair the plan. While some approaches exist [1, 2], none has yet exploited the semantics of the task and the actions. As the actions are meant to be applied in the real world, the meaning of the action is important and may be used not only to repair plans and post-process them, but also to explain them to users. Moreover, plans involving certain actions and tasks that are not accompanied by a real-world context cannot be guaranteed to be safe or trustworthy for users. While a full specification of task and action semantics is cumbersome due to the size and complexity of open domains, some ongoing efforts are addressing the also general, but more manageable domain of common sense. For example, OpenCyc has been running for decades to "assemble a comprehensive ontology and knowledge base that spans the basic concepts and rules about how the world works" [3]. More recently, the knowledge graph community has advanced ground in integrating various knowledge bases (e.g. ATOMIC, ConceptNet, FrameNet, Roget, Visual Genome, Wikidata, WordNet) of common-sense knowledge, in a hyper-relational graph called Common Sense Knowledge Graph [4] (CSKG). A large number of the symbolic representations (e.g. concepts, relations, rules, etc.) in CSKG are relevant for, and could be used as semantic descriptions of, tasks and actions in robot planning. In this project, we propose to combine the ideas of Symbolic Planning for decision-making in robotics with explicit representations of common-sense knowledge in knowledge graphs for safer planning. The idea is that such a combination can leverage contextual descriptions of domains and use common-sense reasoning to avoid plans containing actions or tasks with the potential of being unsafe or untrustworthy. Furthermore, this may also allow to not only improve plans that might not be executable due to semantic constraints, but also to change them in a way that enhances user trust in the robotic system. The symbolic nature of common-sense knowledge graphs such as the CSKG can provide a layer of explainability ensuring that plans can be understood and debugged by humans, creating feedback loops between the planner and the knowledge graph. More specifically, this project:

- Assesses common-sense knowledge in explicit symbolic representations, such as those provided by CSKG and other related datasets, as reliable sources of semantic information for robot planning.
- Develops new planning algorithms that leverage common-sense knowledge graphs and common-sense reasoning to propose semantically rich, explainable plans for robots.
- Evaluates the performance, safety and trustworthiness of these implementations by comparing them with existing approaches that do not exploit common sense.

References

[1] Bercher, P., Biundo, S., Geier, T., Hoernle, T., Nothdurft, F., Richter, F., & Schattenberg, B. (2014, May). Plan, repair, execute, explain—how planning helps to assemble your home theater. In Proceedings of the International Conference on Automated Planning and Scheduling (Vol. 24, pp. 386-394).
[2] Fox, M., Gerevini, A., Long, D., & Serina, I. (2006, June). Plan Stability: Replanning versus Plan Repair. In ICAPs (Vol. 6, pp. 212-221).
[3] Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11), 33-38.
[4] Ilievski, F., Szekely, P., & Zhang, B. (2021). Cskg: The commonsense knowledge graph. In The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6—10, 2021, Proceedings 18 (pp. 680-696). Springer International Publishing.

Adaptation and effective communication in collaborative physically Assistive Tasks

Supervisor: Gerard Canal

Areas: Artificial Intelligence (AI), Human-centred computing

(Back to Scholarship Allocated)

Project Description

Physical robotic Assistance can often be modelled as a collaborative task in which the goal of both the user and the robot is to complete an assistive task together. However, assistive settings have a lot of particularities that differentiate them from traditional Human-Robot Collaboration tasks. For it to be effective, the assistance should be seamless, natural, and without a required effort on the user's side. This means that these robots must be able to communicate with the user in a very natural and intuitive way, but also in an adaptive manner. In this project, we will investigate the development of techniques for the online adaptation of the robot to the human, as well as anticipation of user needs, and seamless communication in the context of assistive tasks such as robotic feeding and dressing.

References

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

Explaining robotic planning decision points along execution

Supervisor: Gerard Canal

Areas: Artificial Intelligence (AI), Human-centred computing

(Back to Scholarship Allocated)

Project Description

Explanation of robotic behaviours has been proved to be very important to improve the understanding of the users of such robots, which improves their trust in the robotic system. However, explanations in robotics are tricky as they need to be given at the correct moment and based on what happened in the execution. In robotic-based planning, an interesting explanation is that of decision points, where the robot could have taken a different action with a different outcome. This project focuses on the explanation of such decision points at execution time, integrating information 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.

Reliability and verification of software for scientific and ML computing

Supervisor: Dr Karine Even-Mendoza, Dr Hana Chockler, Dr Hector Menendez Benito (1st, 2nd and 3rd).

Areas: Machine Learning (ML), Artificial Intelligence (AI), Systems (SE, programming, autonomous systems, robotics, ...), Foundations of computing, Computing Applications

(Back to Scholarship Allocated)

Project Description

Project Description. The long-standing challenge of ensuring the reliability of machine learning implementations (ML) and programming language (PL) libraries, particularly in floating-point and arithmetic computation, has been a focus of attention within programming languages, formal methods, and AI research communities. Historically, researchers have often tended to confine the scope of their research to small-scale problems rather than real-world computer systems. However, modern applications increasingly rely on mathematical computations, such as Alexa voice recognition (using discrete Fourier transform) and autonomous cars. This heightened dependence on accurate computation amplifies concerns about the impact of inaccuracies on system reliability. This project addresses low-level implementations heavily reliant on floating-point computations—a crucial but insufficiently tested area. The precision of these computations significantly influences the reliability of ML and PL libraries and compilers. The project aims to enhance the testing of software systems, specifically focusing on the reliability of software using ML and PL libraries in their algorithms. To achieve this, the student will: - Develop methods to assess the quality of test cases, applying them when attempting to detect miscompilations (silent errors during translation into machine code) and logical bugs in floating-point optimizations and library implementations. - Investigate approaches for testing software libraries and their compilers meaningfully in the context of mathematical and numerical procedures. - Explore fault localization approaches and other techniques to pinpoint detected bugs, ensuring clarity in distinguishing actual bugs from potential issues related to the testing mechanism. The student will employ static and dynamic code analysis, code generation for testing, and testing strategies (like differential testing). After designing a system for meaningful testing of ML and PL libraries, the student will extensively evaluate its bug detection capabilities. The emphasis is on ensuring the identified issues are genuine bugs rather than stemming from the testing methodology. The student will actively engage with the software engineering community, reporting any bugs uncovered during the evaluation process. The above will include investigations of novel ways to design tests and testing campaigns and deal better with coverage of specific functionalities in the compilers and their PL and ML libraries. Context. Compilers and their software libraries, widely used complex programs, are the bridge between software (written in English-look-alike programming language) and machine code (consisting of 0s and 1s). They give us the means to write complex and sophisticated yet efficient algorithms in healthcare, finance, transportation (and more) using mathematical, ML and AI components, empowering today's engineers and relieving them of conceptual high-level tasks. Consequently, compiler bugs broadly impact software, and library defects affect ML and AI trustworthiness. C standard libraries give us the power to compute values of the sinuous function in just one line, and ML libraries allow us to run reinforcement learning with several lines of code. However, ensuring correct translation is complex, as it involves reasoning about the program code's connection to its machine code translation. One of the most expensive yet neglected errors is associated with the floating-point data types: essential types representing numerical data in software, particularly vital in ML and AI implementations; these, in many cases, led to significant financial losses and jeopardised lives. Yet, testing support is often insufficient, commonly limited to the detection of logical faults in lines of code written by the user or crashes when executing machine code because of the testing mathematical code complexity.

References

[1] K. Even-Mendoza, A. Sharma, A. F. Donaldson, and C. Cadar. 2023. GrayC: Greybox Fuzzing of Compilers and Analysers for C. ISSTA 2023: 1219—1231. https://doi.org/10.1145/3597926.3598130

[2] K. Even-Mendoza, C. Cadar, and A. F. Donaldson, CSMITHEDGE: more effective compiler testing by handling undefined behaviour less conservatively. Empir Software Eng 27, 129 (2022). https://doi.org/10.1007/s10664-022-10146-1.

[3] MLighter is an ongoing project with a webpage: http://mlighter.freedevelop.org

[4] K. Even-Mendoza, A. E. J. Hyvarinen, H. Chockler and N. Sharygina: Lattice-based SMT for Program Verification. MEMOCODE 2019 : 16:1-16:11.

[5] H. Chockler, K. Even, and E. Yahav. Finding rare numerical stability errors in concurrent computations. ISSTA, 2013, pages 12—22. (alphabetic order)

[6] J. M. Zhang, M. Harman, L. Ma and Y. Liu, "Machine Learning Testing: Survey, Landscapes and Horizons", in IEEE

Transactions on Software Engineering, vol. 48, no. 1, pp. 1-36, 1 Jan. 2022, doi: 10.1109/TSE.2019.2962027.

[7] J. Chen, J. Patra, M. Pradel, Y. Xiong, H. Zhang, D. Hao, and L. Zhang. 2020. A Survey of Compiler Testing. ACM Comput. Surv. 53, 1, Article 4 (January 2021), 36 pages. https://doi.org/10.1145/3363562

[8] X. Yang, Y. Chen, E. Eide, and J. Regehr. 2011. Finding and understanding bugs in C compilers. SIGPLAN Not. 46, 6 (June 2011), 283—294. https://doi.org/10.1145/1993316.1993532

[9] A. F. Donaldson, H. Evrard, and P. Thomson. 2020. Putting randomized compiler testing into production (experience report). In Proceedings of the 34th European Conference on Object-Oriented Programming (ECOOP 2020). Schloss Dagstuhl-Leibniz-Zentrum fur Informatik. https://drops.dagstuhl.de/opus/volltexte/2020/13179/

[10] V. Livinskii, D. Babokin, and J. Regehr. 2020. Random testing for C and C++ compilers with YARPGen. Proc. ACM Program. Lang. 4, OOPSLA, Article 196 (November 2020), 25 pages. https://doi.org/10.1145/3428264

[11] C. Murphy, K. Shen, and G. Kaiser. 2009. Automatic system testing of programs without test oracles. In Proceedings of the eighteenth international symposium on Software Testing and Analysis (ISSTA '09). Association for Computing Machinery, New York, NY, USA, 189—200. https://doi.org/10.1145/1572272.1572295

[12] A Google self-driving car caused a crash for the first time. 2016.
https://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report
[13] SGD: commonly used during the training process to update the weights of the neural network based on the gradients of the loss function with respect to the weights. https://keras.io/api/optimizers/sgd/

Towards Robust Reasoning in Large Language Models

Supervisor: Yulan He

Areas: Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP)

(Back to Scholarship Allocated)

Project Description

Context Reasoning is a core aspect of human intelligence, playing a crucial role in tasks such as critical thinking, evaluation and decision-making. With the development of large language models (LLMs), we have witnessed their impressive performance in various natural language processing tasks that involve reasoning processes. For an intelligent system to succeed, it must effectively analyse key information within a given context and provide accurate responses by drawing upon its internal knowledge and available resources. Achieving this is a complex process as LLMs must stay updated with the latest information, remain robust in noisy contexts, and be capable of utilising external tools for verification when necessary.

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

- Enhancing reasoning through tool augmentation based on a neuro-symbolic approach. LLMs could benefit from neuro-symbolic reasoning facilitated by external interpreters, particularly in complex tasks.
- Facilitating model adaptation to reason with the most recent knowledge. This involves model editing and fine-tuning the model with new information while retaining its capacity for reasoning in tasks that it has encountered before.
- Promoting collaboration among multiple agents to facilitate reasoning across diverse domains. When faced with an input from an unfamiliar domain, integrating knowledge from multiple trained agents based on its relevance to the specific input could be advantageous.

References

References:

- 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]

Robotics and Social Justice

Supervisor: Martim Brandao

Areas: Artificial Intelligence (AI), Robotics, Human-centred computing

(Back to Scholarship Allocated)

Project Description

The Responsible Robotics and AI Lab is open to applications for a PhD in blue sky research at the intersection of robotics and social justice. The project sits at the intersection of Computer Science and Social Science, and it is expected that the successful candidate will choose an appropriate co-supervisor at a later stage (this choice will be made in conversation with Martim Brandao, the main supervisor). We are particularly looking for students interested in: - Investigating issues of social justice in robotics (e.g. worker conditions, police misuse, accountability, racism, sexism, colonialism) - Developing new methodologies for anti-[racist/ageist/ableist/sexist/capitalist/colonial] robotics - Abolitionist robotics (e.g. for tackling homelessness, mental health, domestic violence, child welfare and other social problems in humane community-grounded ways, without police involvement) - Robotics for sustainability, robotics and environmental justice The PhD proposal should include a plan suggesting how the chosen factors of social justice will be investigated, which (computer simulation-based) prototypes developed, and how they will be evaluated.

Synthetic video generation: counterfactual explanations

Supervisor: Luis C. Garcia Peraza Herrera

Areas: Artificial Intelligence (AI), Machine Learning (ML), Computer vision

(Back to Scholarship Allocated)

Project Description

Counterfactual explanations provide valuable insights into machine learning models. They reveal the minimum changes required in the input to yield a different output, as illustrated in Fig. 1 below. In the case of deep learning models using images as input [1, 2], the counterfactual explanation is also presented as an image:

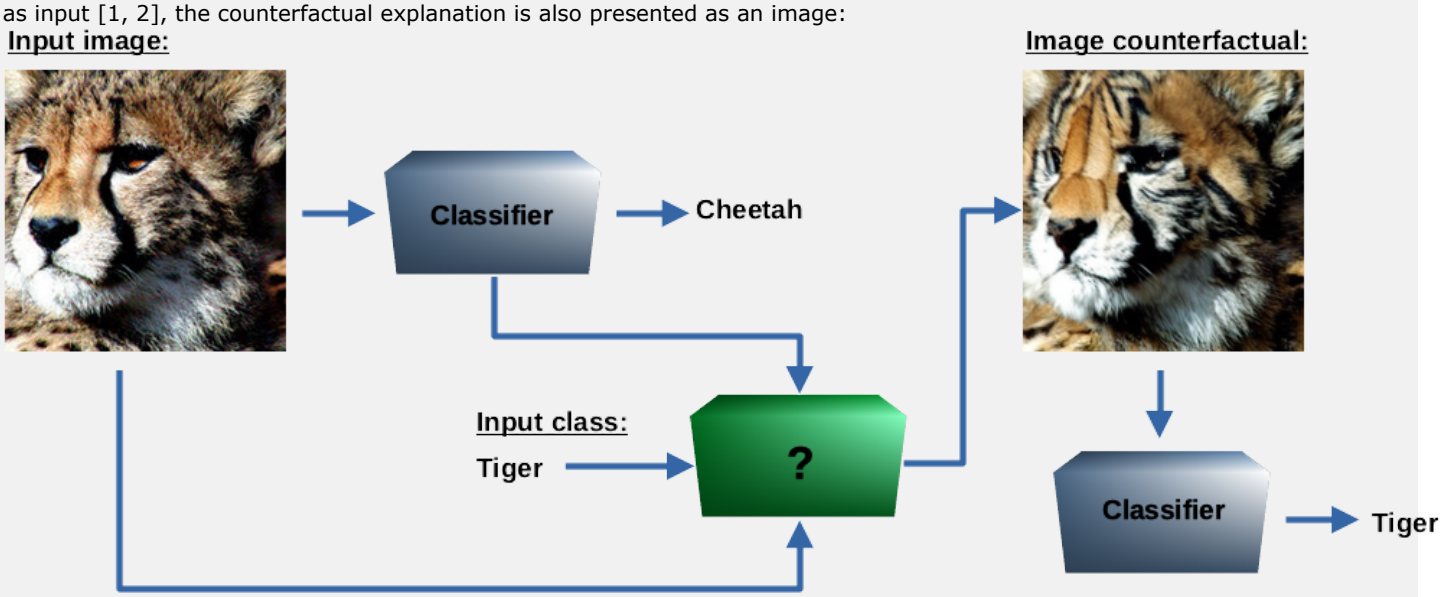

Figure 1. Image counterfactual explanation. The objective of this project is to extend this concept to video data. Instead of dealing with static images, we aim to devise machine learning methods (represented by [?] in Fig. 1 above) for generating video counterfactual explanations. A video counterfactual explanation 1) minimally alters a given input video, and 2) causes the video classifier to predict a different and specific class compared to the original input video. Although our project will focus on developing methods to create video counterfactual explanations specifically tailored to video classifiers, these methods can potentially be applied to other domains as well (e.g. understanding why autonomous robotic systems predict certain actions based on video input). This technology has several applications in the medical domain. Particularly in the realm of computer vision for surgery, the ability to generate synthetic videos has a multitude of potential applications. It holds the capacity to create synthetic datasets for training deep learning models and develop simulators that replicate surgical scenarios, offering clinicians a platform for sharpening their surgical skills.

References

[1] Boreiko et al. Sparse Visual Counterfactual Explanations in Image Space, DAGM GCPR, 2022.
[2] Augustin et al. Diffusion visual counterfactual explanations, NeurIPS, 2022.

Improving Understandability of Automatically Generated Test Cases using Text-to-Text Transformer Models

Supervisor: Gunel Jahangirova

Areas: Systems (SE, programming, autonomous systems, robotics, ...), Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning (ML)

(Back to Scholarship Allocated)

Project Description

The costs associated with software testing activities make their full automation an important research topic. The existing automated test case generation tools (ATGTs) have made significant progress in achieving high coverage, high fault detection rate and input diversity. However, the research in software testing is still far from fulfilling its dream of full automation because multiple studies demonstrate that developers find automatically generated test cases hard to read and understand. This project proposes three directions to tackle the problem of the understandability of automatically generated test cases. The first direction is based on the insight that developer-written test suites capture the information about what testing the given class looks like when performed by the developer and therefore contains features that make the test cases more understandable. We aim to extract the available understandability-related information from developer-written test suites and transfer it into the automatically generated test cases. Our second direction aims to make the understandability of the test case part of the test case generation process such that it favours the test cases with higher understandability. For this, we want to collect a large dataset with human-annotated understandability scores and train a learning model that can predict the understandability score for a candidate test case. The last direction aims to take advantage of the increasing success of text-to-text transformer models. We plan to collect a large dataset of pairs of automatically generated and developer-written test cases that test similar behaviour and train a transformer model that takes an automatically generated test case and transforms it into a version that looks like developer-written. The expected results from the project are in two directions. The first one is the deepened comprehension of the understandability problem. The second one is the set of automated software testing tools that will produce an output that is more understandable by the developers leading to wider adoption of such tools in industrial settings. Moreover, we plan to conduct large studies involving human participants to evaluate the understandability, which will hopefully provide the software engineering research community with examples of well-designed studies evaluating the qualitative properties of test cases.

References

Related Work:

1. E. Daka, J. M. Rojas, and G. Fraser, “Generating unit tests with descriptive names or: Would you name your children thing1 and thing2?” in Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis, 2017, pp. 57–67.
2. G. Fraser, M. Staats, P. McMinn, A. Arcuri, and F. Padberg, “Does automated white-box test generation really help software testers?” in Proceedings of the 2013 International Symposium on Software Testing and Analysis, 2013, pp. 291–301.
3. J. M. Rojas, G. Fraser, and A. Arcuri, “Automated unit test generation during software development: A controlled experiment and think-aloud observations,” in Proceedings of the 2015 international symposium on software testing and analysis, 2015, pp. 338–349.
4. E. Daka, J. Campos, G. Fraser, J. Dorn, and W. Weimer, “Modeling readability to improve unit tests,” in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, 2015, pp. 107–118.

Planning and Reinforcement Learning for versatile autonomous robots

Supervisor: Matteo Leonetti

Areas: Machine Learning (ML), Artificial Intelligence (AI)

(Back to Scholarship Allocated)

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. 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.
2. Meta-Reinforcement Learning for Heuristic Planning. Ricardo Luna Gutierrez, Matteo Leonetti. Proc. of International Conference on Automated Planning and Scheduling (ICAPS), 2021.
3. A synthesis of automated planning and reinforcement learning for efficient, robust decision-making. Matteo Leonetti, Luca Iocchi, Peter Stone. Artificial Intelligence, 2016.
4. Planning in action language BC while learning action costs for mobile robots. Piyush Khandelwal, Fangkai Yang, Matteo Leonetti, Vladimir Lifschitz, Peter Stone. Proc. of International Conference on Automated Planning and Scheduling (ICAPS), 2014.

Safe Reinforcement Learning from Human Feedback

Supervisor: Yali Du

Areas: Artificial Intelligence (AI), Machine Learning (ML), Human-centred computing

(Back to Scholarship Allocated)

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] 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. To address these challenges, this research will leverage the principles of the Abstract Interpretation framework [3], a theory that dictates how to obtain sound, computable, and precise finite approximations of potentially infinite sets of behaviours. Based on the abstraction of states, we aim to enable agents to build a knowledge base for (un)safe behaviours, and thus construct a scheme for when to actively query humans. Besides, due to the nature of sequential decision making, this project will consider temporal abstractions of behaviours and feedback to improve the consistency in safety control. Furthermore, by the effective abstractions, we aim to make the neural-network based agents invariant to task-irrelevant details, and thus generalizable to new downstream tasks.

References

[1] Ilias Kazantzidis, Tim Norman, Yali Du, Christopher Freeman. How to train your agent: Active learning from human preferences and justifications in safety-critical environments. AAMAS 2022.
[2] Runze Liu, Fengshuo Bai, Yali Du, Yaodong Yang. Meta-Reward-Net: Implicitly Differentiable Reward Learning for Preference-based Reinforcement Learning. NeurIPS 2022.
[3] Cousot, P. and Cousot, R. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In Symposium on Principles of Programming Languages (POPL), 1977.

Unifying Principals in Safe and Trusted Assistive AI

Supervisor: Yali Du

Areas: Artificial Intelligence (AI), Machine Learning (ML)

(Back to Scholarship Allocated)

Project Description

AI agents are often required to assist humans in many day-to-day tasks, such as in recommendation systems, restaurant reservation and self-driving cars [1]. As AI agents are frequently evaluated in terms of performance measures, such as human-stated rewards, many challenges are posed. Firstly, due to the involvement of multiple users, agents have to learn to strike a balance between the widely different human preferences [3]. Secondly, while it is usually assumed that humans are acting honestly in specifying their preference, such as by rewards or demonstrations, the consequence of humans mis-stating their objectives is commonly underestimated. Humans may maliciously or unintentionally mis-state their preference, leading the assistive AI agent to perform unexpected implementations. An example is the Tay chatbot from Microsoft; prankster users falsify their demonstrations and train Tay to mix the racist comments into its dialogue . This project aims to unify many principals to achieve fairness and social welfare, towards building safe and trustworthy assistive AI agents that avoid bias and manipulation like Tay Chatbot. The human preference can be explicitly stated as 'like' or 'dislike' of the agent's performance, or implicitly stated through the demonstrations. Two popular learning paradigms can be considered, Reinforcement Learning (RL) from specified preference [1] and Apprenticeship Learning (AL) [2] with human's value implicitly expressed by their demonstrations. By reinforcement learning, agents learn to perform given tasks based on preference. By apprenticeship learning, agents observe human demonstrations (historical trajectories) that reveal human's interest, and learn to perform tasks to align with human values. Example questions that can be explored: Multi-objective learning: given the objectives specified either by reward or demonstrations, how can we balance the different and possibly conflicting objectives from users? Manipulating the assistive learning: a famous result from social choice theory is that, a non-trivial collective decision is subject to manipulation [4], how easy is it for one or some users to change the behavior of an assistive agent? Or how can a human bias the system towards their own interest? By studying how to manipulate assistive learning, the ultimate goal is still to develop robots that can delegate multiple humans' interests fairly and correctly.

References

[1] Chen, X., Du, Y., Xia, L., & Wang, J. (2021). Reinforcement recommendation with user multi-aspect preference. The Web Conference 2021 - Proceedings of the World Wide Web Conference (WWW) 2021, 425—435. https://doi.org/10.1145/3442381.3449846
[2] Fickinger, A., Zhuang, S., Critch, A., Hadfield-Menell, D., & Russell, S. (2020). Multi-Principal Assistance Games: Definition and Collegial Mechanisms. NeurIPS, 2020, 1—10.
[3] McAleer S, Lanier J, Dennis M, Baldi P, Fox R. Improving Social Welfare While Preserving Autonomy via a Pareto Mediator. arXiv preprint arXiv:2106.03927. 2021.
[4] Allan Gibbard. Straightforwardness of game forms with lotteries as outcomes. Econometrica: Journal of the Econometric Society, pages 595—614, 1978.

Safe Reinforcement Learning from Human Feedback

Supervisor: Yali Du

Areas: Artificial Intelligence (AI), Machine Learning (ML)

(Back to Scholarship Allocated)

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] 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.

Estimating the ground truth of LLMs in softare engineering Tasks

Supervisor: Jie M. Zhang

Areas: Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), Systems (SE, programming, autonomous systems, robotics, ...)

(Back to Scholarship Allocated)

Project Description

When using LLMs for software engineering tasks such as code generation, it is important to understand how reliable the generated outputs are. Most of the time the ground truth is unknown. Thus, it is important to estimate the confidence and accuracy of LLMs so as to improve their usability and help users judge whether to adopt the provided solutions. This proposal aims to explore different methods to estimate the confidence of LLMs in generating solutions, in particular to software engineering-related tasks.

References

https://arxiv.org/pdf/2310.03533.pdf
https://openreview.net/forum?id=gjeQKFxFpZ

Text promptable surgical video generation

Supervisor: Luis C. Garcia Peraza Herrera

Areas: Artificial Intelligence (AI), Machine Learning (ML), Computer vision

(Back to Scholarship Allocated)

Project Description

The goal of this PhD project is to develop an innovative framework for generating synthetic surgical videos through command prompts. This research aims to advance the field of surgical simulation by creating realistic and diverse datasets for training and evaluating computer vision models in surgery.

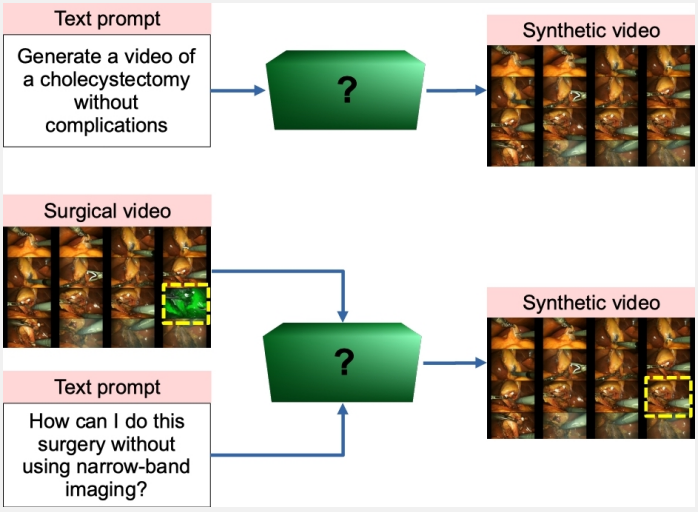

Figure 1. Text promptable surgical video generation. Our ultimate aspiration is to establish a system akin to DALL-E, whereby we can seamlessly request the generation of synthetic surgical videos on-demand. We aim to explore methods that use command prompts as a guiding mechanism, investigating the integration of procedural commands to control the content, complexity, and variability of the simulated surgeries. This involves improving the visual quality, mimicking real-world variations in surgical procedures, and incorporating dynamic factors such as tissue deformation, blood, and instrument interactions. The generated videos should cover a wide spectrum of medical procedures, surgical tools, and operating conditions to improve the robustness and generalization of the simulation.

References

[1] Garcia-Peraza-Herrera et al. Image Compositing for Segmentation of Surgical Tools without Manual Annotations, IEEE TMI, 2021.
[2] Singer et al. Make-A-Video: Text-to-Video Generation without Text-Video Data, ICLR, 2023.

Scholarship Not Allocated

(Back to Top)

- Explaining, improving, and simplifying RL policies using causal reasoning
- Game-theoretic models in cryptoeconomics: incentives, mechanisms and blockchain dynamics
- Learning emergent behaviors in multi-agent systems: game theory and chaos dynamics for artificial intelligence
- Trustworthy digital twins and simulations
- Ensuring Trustworthy AI through Verification and Validation in ML Implementations: Compilers and Libraries via Generative Approaches.
- AI in finance
- Dealing with imperfect rationality in computational systems
- Enhancing Safety in Robotics by Tackling Blind-Spots and Bias in AI Models
- Co-Improving Generative AI Systems
- Data Science Methodologies for Biomedicine
- Improving active learning strategies for limited annotation budgets
- Verification of Autonomous Agents in Uncertain Environments
- Reliable Learning for Safe Autonomy with Conformal Prediction
- Causal Explanations for Sequential Decision Making

Explaining, improving, and simplifying RL policies using causal reasoning

Supervisor: Hana Chockler

Areas: Artificial Intelligence (AI), Machine Learning (ML)

(Back to Scholarship Not Allocated)

Project Description

Reinforcement learning is a powerful method for training policies that complete tasks in complex environments. The policies produced are optimised to maximise the expected reward provided by the environment. While performance is clearly an important goal, the reward typically does not capture the entire range of our preferences. By focusing solely on performance, we risk overlooking the demand for models that are easier to analyse, predict and interpret. The hypothesis of this project proposal is that many trained policies are needlessly complex and suboptimal, i.e., that there exist alternative policies that perform just as well or even better while being significantly simpler. Furthermore, these policies can be extracted from a given trained policy using causal reasoning. Moreover, causal analysis can help to extract policy explanations, which are small subsets of policy's decisions that are the most important for achieving the reward. Preliminary results were demonstrated in [1,2], and the quantification of causality is presented in [3].

References

[1] Hadrien Pouget, Hana Chockler, Youcheng Sun, Daniel Kroening:
Ranking Policy Decisions. NeurIPS 2021: 8702-8713.
[2] Daniel C. McNamee, Hana Chockler: Causal policy ranking. CoRR abs/2111.08415 (2021).
[3] Hana Chockler, Joseph Y. Halpern: Responsibility and Blame: A Structural-Model Approach. J. Artif. Intell. Res. 22: 93-115 (2004).

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

Supervisor: Stefanos Leonardos

Areas: Machine Learning (ML), Artificial Intelligence (AI), Foundations of computing

(Back to Scholarship Not Allocated)

Project Description

This project targets students interested in advancing cutting-edge research at the intersection of game theory and cryptoeconomics. The project's aim is to model and analyze blockchain-enabled economies through a game-theoretic lens. Special focus will be placed on transaction fee markets, miner extractable value (MEV) incentives, proposer-builder separation in Ethereum block creation, MEV-boost auctions, transaction censorship, attacks in decentralized exchanges, and related phenomena. The study will explore cryptoeconomic mechanisms, dissecting participant 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, 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
3. 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
4. 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
5. 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.
6. 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.
7. 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.

Learning emergent behaviors in multi-agent systems: game theory and chaos dynamics for artificial intelligence

Supervisor: Stefanos Leonardos

Areas: Artificial Intelligence (AI), Machine Learning (ML), Foundations of computing

(Back to Scholarship Not Allocated)

Project Description

This project targets 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. In this context, 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. 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.
2. 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/dd1970fb03877a235d530476eb727dab-Paper.pdf.
3. 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.
4. 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
5. Cheung, Y.K., Leonardos, S., and Piliouras, G. (2021). Learning in Markets: Greed Leads to Chaos but Following the Price is Right. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Main Track, pp. 111-117, doi:10.24963/ijcai.2021/16.

Trustworthy digital twins and simulations

Supervisor: Steffen Zschaler

Areas: Artificial Intelligence (AI), Systems (SE, programming, autonomous systems, robotics, ...)

(Back to Scholarship Not Allocated)

Project Description

Digital twins are digital representations of real-world entities that are continually updated and can affect changes in the real world. They are used as decision-support systems or even to make decisions autonomously. It is, therefore paramount that we can trust them and have a clear and explicit understanding of when they are applicable or not. However, the scope of validity and rationale underpinning a digital twin is often not explicitly documented or is only documented in natural-language text, making evaluating and maintaining any such arguments difficult. In this project, you will explore how trust in digital twins can be better supported through developing structured, explicitly represented arguments and how these arguments can be semi-automatically validated and maintained over time. I work with stakeholders in a range of domains, providing opportunities for research in different contexts, potentially as separate PhD projects.

Ensuring Trustworthy AI through Verification and Validation in ML Implementations: Compilers and Libraries via Generative Approaches.

Supervisor: Karine Even-Mendoza, Hector Menendez Benito

Areas: Machine Learning (ML), Systems (SE, programming, autonomous systems, robotics, ...), Artificial Intelligence (AI), Computing Applications

(Back to Scholarship Not Allocated)

Project Description

Project Description. The issue of machine learning trust is a pressing concern that has brought together multiple communities to tackle it. With the increasing use of tools such as ChatGPT and the identification of fairness issues, detecting security concerns and ensuring the reliability of machine learning is paramount to its continued development. This project addresses low-level implementation in machine learning, an often-overlooked area, but one that profoundly impacts the reliability of libraries and languages, including TensorFlow, Keras, PyTorch, Python, and R. Knowledge in programming languages and compilers such as CPython and C, as well as familiarity with ML libraries in Python and R, are essential for this project. Project. The project's main idea is to generate and diversify test cases for testing machine learning implementations for each level of abstraction from the top language to the low-level libraries. The student will be employing diverse testing techniques, like LLM for generating test cases, focusing on aspects like numerical validity, security, and fairness to be able to test these aspects more thoroughly, and a "Godel Test" variant: a method that parametrises input generators for programs and controls the parameters to create testing strategies. Among the testing strategies, we will apply multiple test suite generation strategies, such as focused testing (i.e. testing new software's components, which are common in the traditional machine learning libraries), vulnerability unmasking, and differential testing techniques. The student will design a system based on search strategies that will try to guide the algorithms to exhibit the possible branches of the machine learning code and its compilers. For that, we will extend the testing framework of the MLighter tool, a holistic tool for evaluating the security, reliability and performance of machine learning, to deal with these specific problems using generative approaches (including LLM). The student will then extensively evaluate the system's capabilities, focusing on its ability to test deeper parts of the ML code and potentially communicating with the software engineering community to report any exposed vulnerabilities and logical bugs discovered during the evaluation process. The above will include investigations of novel ways to design tests and testing campaigns using LLM and better deal with coverage of specific functionality in the ML code and the test oracle problem. Context. To the best of our knowledge, while there are a few works related to Python compiler fuzzing (and none for R compiler fuzzing), we have recently seen a substantial volume of research focused on testing ML libraries. With the introduction of LLM and the growing interest in ChatGPT-related research, there is an increased need to expand and enhance testing methodologies in these areas, including a growing emphasis on fuzzing ML libraries. None of these works suggested a holistic way of dealing with the reliability of machine learning libraries with the compilers generating their executable binaries. Considering the potential points of failure: it can occur in any of the following components, or a combination thereof: (1) the Python or R compiler, (2) the ML library implemented in an optimising compiler like C, and (3) the optimising compiler itself (e.g., C). The project consists of two parts: ML libraries testing and the lowest level of testing, which is compiler testing.

References

[1] MLighter is an on-going project with a webpage: http://mlighter.freedevelop.org

[2] H. D. Menendez, "Measuring Machine Learning Robustness in front of Static and Dynamic Adversaries*," 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), Macao, China, 2022, pp. 174-181, doi: 10.1109/ICTAI56018.2022.00033.

[3] K. Even-Mendoza, A. Sharma, A. F. Donaldson, and C. Cadar. 2023. GrayC: Greybox Fuzzing of Compilers and Analysers for C. In Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2023). Association for Computing Machinery, New York, NY, USA, 1219—1231. https://doi.org/10.1145/3597926.3598130

[4] E. T. Barr, M. Harman, P. McMinn, M. Shahbaz and S. Yoo, "The Oracle Problem in Software Testing: A Survey," in IEEE Transactions on Software Engineering, vol. 41, no. 5, pp. 507-525, 1 May 2015, doi: 10.1109/TSE.2014.2372785.

[5] A. Wei, Y. Deng, C. Yang, and L. Zhang. 2022. Free lunch for testing: fuzzing deep-learning libraries from open source. In Proceedings of the 44th International Conference on Software Engineering (ICSE '22). Association for Computing Machinery, New York, NY, USA, 995—1007. https://doi.org/10.1145/3510003.3510041

[6] O. Bastani, R. Sharma, A. Aiken, and P. Liang. 2017. Synthesizing program input grammars. SIGPLAN Not. 52, 6 (June 2017), 95—110. https://doi.org/10.1145/3140587.3062349

[7] CompCert: Leroy, X. (2021). The CompCert C verified compiler: Documentation and user's manual (Doctoral dissertation, Inria).

[8] S. Poulding and R. Feldt. 2014. Generating structured test data with specific properties using nested Monte-Carlo search. In Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation (GECCO '14). Association for Computing Machinery, New York, NY, USA, 1279—1286. https://doi.org/10.1145/2576768.2598339

[8] J. Chen, J. Patra, M. Pradel, Y. Xiong, H. Zhang, D. Hao, and L. Zhang. 2020. A Survey of Compiler Testing. ACM Comput. Surv. 53, 1, Article 4 (January 2021), 36 pages. https://doi.org/10.1145/3363562

[9] A. Dakhama, K. Even-Mendoza, W.B. Langdon, H. Menendez, and J. Petke (2023). SearchGEM5: Towards Reliable gem5 with Search Based Software Testing and Large Language Models. Symposium on Search Based Software Engineering (SSBSE). https://tinyurl.com/2u2aeb4r

[10] J. M. Zhang, M. Harman, L. Ma and Y. Liu, "Machine Learning Testing: Survey, Landscapes and Horizons", in IEEE Transactions on Software Engineering, vol. 48, no. 1, pp. 1-36, 1 Jan. 2022, doi: 10.1109/TSE.2019.2962027.

[11] The Ken Thompson Heck: https://wiki.c2.com/?TheKenThompsonHack

AI in finance

Supervisor: Carmine Ventre

Areas: Artificial Intelligence (AI)

(Back to Scholarship Not Allocated)

Project Description

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/.

Dealing with imperfect rationality in computational systems

Supervisor: Carmine Ventre

Areas: Foundations of computing, Artificial Intelligence (AI)

(Back to Scholarship Not Allocated)

Project Description

Computational/AI systems often collect their input from humans. For example, parents are asked to input their preferences over primary schools before a centralised algorithm allocates children to schools. Should the AI trust the input provided by parents who may try to game the system? Should the parents trust that the AI system has optimised their interests? Would it be safe to run the algorithm with a potentially misleading input? Algorithmic Game Theory (AGT) is a research field that attempts to add safety and trustworthiness to AI systems vis-a-vis strategic reasoning. With its set of symbolic tools, one aims to align the goals of the AI system (e.g., the allocation algorithm above) with those of the agents (e.g., the parents above) involved. The AI will then be safe, in that we can analytically predict end states of the system, and trustworthy, since no rational agent will attempt to misguide the system and the system will work on truthful inputs. One assumption underlying much of the work in AGT is, however, pretty limiting: agents need to be fully rational. This is unrealistic in many real-life scenarios; we, in fact, have empirical evidence that people often misunderstand the incentives and try to game the system even when it is against their own interest. Moreover, modern software agents, often built on top of AI tools, are seldom able to perfectly optimise their rewards. This project will look at novel approaches to deal with imperfect rationality, including analysis of known AI systems and the design of novel ones. This will involve theoretical work that builds on the recent advances on mechanism design for imperfectly rational agents (namely obvious strategyproofness and not obvious manipulability) to include more complex domains and the modelling of further behavioural biases in mechanism design. Prospective applicants are encouraged to consult the publications of Prof Ventre at https://kclpure.kcl.ac.uk/portal/en/persons/carmine.ventre/publications/.

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

Supervisor: Gerard Canal (1st) and Hector Menendez (2nd)

Areas: Machine Learning (ML), Artificial Intelligence (AI), Systems (SE, programming, autonomous systems, robotics, ...)

(Back to Scholarship Not Allocated)

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.

Co-Improving Generative AI Systems

Supervisor: Dr Hector Menendez Benito (1st) and Dr Karine Even-Mendoza (2nd)

Areas: Artificial Intelligence (AI), Machine Learning (ML), Computer vision, Data science

(Back to Scholarship Not Allocated)

Project Description

Generative AI is non-deterministic; therefore, creating an appropriate test suite to evaluate whether a generative AI is reaching its expected outcome becomes a difficult task. Thus, it can hardly be measurable with metrics currently used to evaluate the quality of software. In the context of generative AI, we need to consider that for every prompt the user introduces, a different outcome can be created. Consequently, multiple repetitions of the prompt are required to understand whether the learning process has been fulfilled. It also applies to the different parameters of the models. In many cases, this requires manual evaluation, which is difficult and costly to scale. In the context of image generation, this problem can be alleviated under specific constraints by using a combination of different AIs. In our previous work, we created a tool called StableYolo that was able to select proper parameters for the generative AI process (under the Stable Diffusion model) by using automatic feedback from a visualization model (YOLO). This automatic feedback was focused mainly on photorealistic images, and in combination with search, it was able to identify proper parameters for the system and engineer both the positive and negative prompts to the best possible combination of words. This PhD proposal aims to focus on the generalization problem of this strategy. The main goal is to investigate how different artificial intelligence models can be combined to improve their quality. The student will start by extending the idea of generative AI in images, focusing not only on a photorealistic environment but also on other possible environments. This will also attempt to generate multiple objectives for the optimization process that aim to improve not only the quality of the problems but also to identify new words and combinations of AIs to support the description process. The main idea is to create a general framework to support how AIs should be combined to reinforce each other. The project will be divided into three parts as follows. First part: model identification and matching. During the first part, the student will focus on studying the state of the art regarding different models for generative AI. Within this model, the student will try to match which ones should support each other. In a similar fashion, the student will identify the parameters of the models and study how these parameters affect the output's quality. With this information, the student will be able to create a search algorithm that co-evolves and involves both models. Second part: formal auditing of the generative model. This part focuses on creating or identifying different metrics to measure the effect of the optimization process, define boundaries during the optimization, and create a new set of strategies that will support identifying other kinds of problems within the systems, for instance, bias or fairness issues. Third part: improving and explaining the models. The last part focuses on how the models can be directly improved and not only turned based on the outputs of the other models. The end goal of this strategy is to create better AI systems with a focus on adversarial machine learning combined with explainability.

References

- H. Berger, A. Dakhama, Z. Ding, K. Even-Mendoza, D. Kelly, H. D. Menendez, R. Moussa, and F. Sarro, StableYolo: Optimizing Image Generation for Large Language Models. In Symposium on Search-Based Software Engineering (SSBSE) 2023 Springer.
- X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, H. Wang. Large language models for software engineering: A systematic literature review. arXiv preprint arXiv:2308.10620. 2023 Aug 21
- J. M. Zhang, M. Harman, L. Ma and Y. Liu, "Machine Learning Testing: Survey, Landscapes and Horizons", in IEEE Transactions on Software Engineering, vol. 48, no. 1, pp. 1-36, 1 Jan. 2022, doi: 10.1109/TSE.2019.2962027.
- Jo, A. (2023). The promise and peril of generative AI. Nature, 614(1), 214-216.
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35, 36479-36494.

Data Science Methodologies for Biomedicine

Supervisor: Dr Sophia Tsoka

Areas: Artificial Intelligence (AI), Machine Learning (ML), Data science

(Back to Scholarship Not Allocated)

Project Description

Data representation and modelling via supervised and unsupervised learning are key procedures in Biomedical Data Science. Next-generation sequencing (NGS) data - in some cases spatially-resolved or at single-cell resolution - are revolutionising data mining approaches for complex diseases and necessitate deployment of machine learning models able to handle large, sparse and heterogeneous inputs. This project addresses the development and application of machine learning (ML) methods to model data and signalling interactions in complex diseases, and link them to response outcomes such as prognosis or treatment effects [1]. Specific tasks are outlined, but not limited to, below.

Data integration and representation. The task of data integration and management in the context of clinical applications is particularly challenging due to complex data of heterogeneous nature, i.e. arising from diverse sources of measurement and requiring different types of processing. Graph databases and knowledge graphs (KGs) can be used to model such integration, for example towards drug repurposing tasks [2, 3]. Representation, integration and link prediction tasks will be implemented via KGs across biomedical data domains (imaging, clinical and NGS data) for exploration, visualisation and analysis.

Unsupervised learning through network analysis. In clustering algorithms, the integration of multiple data sources in multilayer networks to detect composite clusters captures molecular functions more accurately. We reported combinatorial optimisation methods for consensus clustering [4], to model multiplex networks, determine composite communities and link topological to functional properties. Extensions of this work are envisaged for application on multiple NGS data types, to include meta-data annotations and to rationalise the choice of data layers in the model through information theory.

Informed machine learning models. Typically, disease classification tasks assume that model features are independent. Incorporating a priori knowledge of relations between predictors (for example through known interaction events) can decrease data dimensionality and increase biological interpretability of ML models. Previously we reported the development of mathematical optimisation models for pathway activity inference applied in cancer subtype prediction [5]. We will extend such modelling frameworks to incorporate prior-knowledge in neural networks, to enhance interpretability of deep learning models and specify robust predictor signatures.

References

References

1. E. Amiri Souri, A. Chenoweth, A. Cheung, S.N. Karagiannis, S. Tsoka. Cancer Grade Model: a multi-gene machine learning-based risk classification for improving prognosis in breast cancer. Br J Cancer, 125, 748-758, 2021.

2. E. Amiri Souri, R. Laddach, S.N. Karagiannis, L.G. Papageorgiou, S. Tsoka. Novel drug-target interactions via link prediction and network embedding. BMC Bioinformatics, 23, 121, 2022.

3. E. Amiri Souri, A. Chenoweth, S.N. Karagiannis, S. Tsoka. Drug repurposing and prediction of multiple interaction types via graph embedding. BMC Bioinformatics, 24, 202, 2023.

4. L. Bennett, A. Kittas, G. Muirhead, L. G. Papageorgiou, S. Tsoka. Detection of composite communities in multiplex biological networks, Scientific Reports, 5, 10345, 2015.

5. Y. Chen, S. Liu, L.G. Papageorgiou, K. Theofilatos, S. Tsoka. Optimisation Models for Pathway Activity Inference in Cancer, Cancers, 15, 1787, 2023.

Improving active learning strategies for limited annotation budgets

Supervisor: Luis Carlos Garcia-Peraza

Areas: Artificial Intelligence (AI), Machine Learning (ML), Computer vision

(Back to Scholarship Not Allocated)

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. Active learning problem Figure 1. The active learning problem. 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

[1] Mahmood et al. Low-Budget Active Learning via Wasserstein Distance: An Integer Programming Approach, ICLR, 2022.
[2] Chen et al. Making Your First Choice: To Address Cold Start Problem in Medical Active Learning, MIDL, 2023.

Verification of Autonomous Agents in Uncertain Environments

Supervisor: Nicola Paoletti

Areas: Machine Learning (ML), Artificial Intelligence (AI), Systems (SE, programming, autonomous systems, robotics, ...)

(Back to Scholarship Not Allocated)

Project Description

With the widespread deployment of autonomous agents, such as autonomous cars and robots and the increasing focus on AI safety, this project aims to investigate the safety of neuro-symbolic agents. The field of neuro-symbolic systems is an exciting area of research that combines the power of machine learning with the rigour of symbolic reasoning. Neural systems have shown great promise in a wide range of applications, from robotics and autonomous systems to natural language processing and decision-making. However, verifying the correctness of these systems remains a significant challenge. While neural networks are excellent at learning patterns in data, they can be difficult to interpret and analyse. On the other hand, symbolic reasoning is highly transparent and understandable, but it can be challenging to scale up to complex non-linear and high-dimensional systems. In this project, we are interested in the analysis of multi-agent neuro-symbolic systems (NSS), which are systems comprising multiple agents interacting with each other and with the environment. The behaviour of such agents is determined by a combination of physical dynamics, such as laws of motion, and machine learning components, which are used, for instance, for perception and control. This kind of systems is relevant in many applications, such as multi-agent (deep) reinforcement learning [1], swarm robotics, and traffic management. We aim to develop verification algorithms for multi-agent NSSs, to provide formal guarantees about the satisfaction of some requirements of interest (reach-avoid, multi-stage tasks, or other kinds of temporal properties). Formal reasoning about these systems is, however, computationally challenging, owing to the presence of (complex) neural network models, multiple agents, uncertain (non-deterministic or probabilistic) environments, and sequential decision-making over multiple time steps. Considerable progress has been made in the verification of one-step reachability for neural networks (i.e., input-output specifications), including probabilistic deep models, using techniques like bound propagation [2,3], constraint solving [4,5], and abstract interpretation [6]. These techniques have been recently extended to the verification of single-agent sequential decision-making [7-9]. However, the multi-agent case remains a largely unexplored research area, with the exception of [10-12]. This project will focus on developing new methods to verify the behaviour of multi-agent NSSs under uncertain environments, where uncertainty can be reasoned about in a probabilistic or non-deterministic fashion. We envision that the solution methods will build on and improve existing verification techniques for single-agent systems, possibly investigating suitable abstractions for dimensionality reduction as well as the combination with data-driven methods like [13] to obtain probabilistic guarantees in the most complex cases where purely symbolic approaches fail. The research project will contribute to the development of trustworthy and reliable multi-agent systems, which can have a significant impact on many applications. The proposed techniques will be evaluated in standard multi-agent RL benchmarks like [14] and different real-world scenarios coming from the REXASI-PRO EU project [15], which will focus on safe navigation of autonomous wheelchairs in crowded environments for people with reduced mobility.

References

[1] Hernandez-Leal, Pablo, Bilal Kartal, and Matthew E. Taylor. "A survey and critique of multiagent deep reinforcement learning." Autonomous Agents and Multi-Agent Systems 33, no. 6 (2019): 750-797.
[2] Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., ... & Kohli, P. (2018). On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint arXiv:1810.12715.
[3] Wicker, Matthew, Luca Laurenti, Andrea Patane, and Marta Kwiatkowska. "Probabilistic safety for Bayesian neural networks." In Conference on uncertainty in artificial intelligence, pp. 1198-1207. PMLR, 2020.
[4] Katz, Guy, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah et al. "The marabou framework for verification and analysis of deep neural networks." In CAV 2019, pp. 443-452.
[5] Botoeva, E., Kouvaros, P., Kronqvist, J., Lomuscio, A., & Misener, R. (2020, April). Efficient verification of relu-based neural networks via dependency analysis. In AAAI Conference on Artificial Intelligence, AAAI.
[6] Singh, G., Gehr, T., Puschel, M., & Vechev, M. (2019). An abstract domain for certifying neural networks. Proceedings of the ACM on Programming Languages, POPL.
[7] Lopez, Diego Manzanas, et al. "NNV 2.0: the neural network verification tool." International Conference on Computer Aided Verification. Cham: Springer Nature Switzerland, 2023.
[8] Wicker, Matthew, et al. "Probabilistic Reach-Avoid for Bayesian Neural Networks." arXiv preprint arXiv:2310.01951 (2023).
[9] Hosseini, M. & Lomuscio, M., (2023) Bounded and Unbounded Verification of RNN-based Agents in Non-deterministic Environments. In AAMAS 2023.
[10] Akintunde, Michael E., Elena Botoeva, Panagiotis Kouvaros, and Alessio Lomuscio. "Verifying strategic abilities of neural-

symbolic multi-agent systems." In Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, vol. 17, no. 1, pp. 22-32. 2020.
[11] Mqirmi, P. E., Belardinelli, F., & Leon, B. G. (2021). An Abstraction-based Method to Check Multi-Agent Deep Reinforcement-Learning Behaviors. In The International Conference on Autonomous Agents and Multiagent Systems, AAMAS.
[12] Yan, Rui, Gabriel Santos, Gethin Norman, David Parker, and Marta Kwiatkowska. "Strategy synthesis for zero-sum neuro-symbolic concurrent stochastic games." arXiv preprint arXiv:2202.06255 (2022).
[13] Bortolussi, Luca, Francesca Cairoli, and Nicola Paoletti. "Conformal Quantitative Predictive Monitoring of STL Requirements for Stochastic Processes." In 26th ACM International Conference on Hybrid Systems: Computation and Control. 2023.
[14] Mordatch, Igor, and Pieter Abbeel. "Emergence of grounded compositional language in multi-agent populations." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.
[15] REliable & eXplAinable Swarm Intelligence for People with Reduced mObility (REXASI-PRO), https://cordis.europa.eu/project/id/101070028.

Reliable Learning for Safe Autonomy with Conformal Prediction

Supervisor: Nicola Paoletti

Areas: Artificial Intelligence (AI), Machine Learning (ML), Systems (SE, programming, autonomous systems, robotics, ...)

(Back to Scholarship Not Allocated)

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 model-driven 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://rexasi-pro.spindoxlabs.com/.

Causal Explanations for Sequential Decision Making

Supervisor: Nicola Paoletti

Areas: Artificial Intelligence (AI)

(Back to Scholarship Not Allocated)

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.