PhD projects (Foundations of computing)

- Foundations of Hybrid Search
- Benevolent Control: Algorithmic Foundations and Dynamic Modelling of Affirmative Interventions in Collective Decision-Making
- Strategic and Structural Robustness of LLM-Mediated Social Choice: Manipulation, Bias, and Control in Generative Democratic Processes
- Computational Foundations of Decentralised Markets
- Mechanism Design for Robust AI Alignment
- Identifying and Shifting the Frontier of Safe Multi-Agent AI
- Biologically Inspired Artificial Intelligence for Understanding and Mitigating Age-Related Cognitive Decline
- Counterfactual Causal Reasoning for Sequential Decision-Making and AI Systems
- Algorithms for perpetual scheduling problems
- Design and analysis of Population Protocols

Foundations of Hybrid Search

Supervisor: Dr Loukides and Dr Charalampopoulos

Project Description

Hybrid search combines keyword-based search with meaning-based (vector) search to give better results than either method alone, and it is now widely used in modern search tools and AI assistants. But many basic questions are still unresolved. For example, the common method for merging the two types of results, called reciprocal rank fusion, is quite sensitive to how it is set up, and simpler weighted-combination methods have actually been shown to work better in many cases. There is also a lack of solid, side-by-side testing of different search methods on real data, so most current advice is based on trial and error rather than clear evidence.

This project will look at three things: how to pick and balance different search methods depending on the type of data; how to automatically tune these systems well without needing lots of manually labeled examples; and how to build fairer, more realistic tests for measuring how well hybrid search actually performs. The project will use ideas from algorithms and data structures, information retrieval, machine learning, and probability-based scoring methods. It will be tested using well-known search benchmarks and real datasets. The expected results are practical guidelines for building better search systems, improved ways of testing them, and open resources that other researchers can use.

References

1. Bruch, S., Gai, S., & Ingber, A. (2024). "An Analysis of Fusion Functions for Hybrid Retrieval." ACM Transactions on Information Systems, 42(1), Article 20:1—20:35. DOI: 10.1145/3596512.
2. Wang, M., Tan, B., Gao, Y., Jin, H., Zhang, Y., Ke, X., Xu, X., & Zhu, Y. (2025). "Balancing the Blend: An Experimental Analysis of Trade-offs in Hybrid Search." CoRR abs/2508.01405.

Benevolent Control: Algorithmic Foundations and Dynamic Modelling of Affirmative Interventions in Collective Decision-Making

Supervisor: Maria Polukarov

Project Description

This project reframes electoral control, traditionally studied as an adversarial threat, as a design tool for improving representation. This idea allows for several alternative approaches. For example, building on a dynamic spatial model where voters and candidates occupy a (weighted) multidimensional metric space that evolves over time, one could (1) formally define and classify ``benevolent'' control operations (e.g., candidate quotas, reserved seats, or shortlist restrictions) alongside existing constructive/destructive control types, and characterise their computational complexity across common voting rules; (2) develop dynamic representation metrics that would extend or replace the traditional justified-representation axioms to remain informative in fragmented electorates where cohesive-group conditions trivially hold; (3) model long-run feedback effects, such as preference and weight adaptation following an intervention, to study when affirmative interventions produce lasting shifts in voter attitudes versus temporary compliance; and (4) empirically validate the model against real-world quota reforms using existing panel data. The thesis will bridge computational social choice, mechanism design, and empirical political economy, producing both hardness/tractability results and practical guidelines for designing interventions that balance robustness against manipulation with capacity to correct structural underrepresentation.

References

John J. Bartholdi, Craig A. Tovey, and Michael A. Trick. 1992. How hard is it to control an election? Mathematical and Computer Modelling 16, 8 (1992), 27—40.

Edith Hemaspaandra, Lane A. Hemaspaandra, and Jorg Rothe. 2007. Anyone but him: The complexity of precluding an alternative. Artificial Intelligence 171, 5-6 (2007), 255—285.

Edith Hemaspaandra, Lane A. Hemaspaandra, and Jorg Rothe. 2017. The complexity of controlling candidate-sequential elections. Theoretical Computer Science 678 (2017), 14 — 21.

Piotr Faliszewski and Jorg Rothe. 2016. Control and Bribery in Voting. In Handbook of Computational Social Choice. 146—168.

Jiehua Chen, Piotr Faliszewski, Rolf Niedermeier, and Nimrod Talmon. 2017. Elections with few voters: candidate control can be easy. J. Artif. Int. Res. 60, 1 (Sept. 2017), 937—1002.

Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. 2017. Justified representation in approval-based committee voting. Social Choice and Welfare 48, 2 (2017), 461—485.

Robert Bredereck, Piotr Faliszewski, Andrzej Kaczmarczyk, and Rolf Niedermeier. 2019. An experimental view on committees providing justified representation. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (Macao, China) (IJCAI'19). AAAI Press, 109—115.

Raghabendra Chattopadhyay and Esther Duflo. 2004. Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. Econometrica 72, 5 (2004), 1409—1443.

Lori Beaman, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. Powerful Women: Does Exposure Reduce Bias? The Quarterly Journal of Economics 124, 4 (2009), 1497—1540.

Martin Lackner and Piotr Skowron. 2023. Multi-Winner Voting with Approval Preferences. Springer. https://doi.org/10.1007/978-3-031-09016-5

Tomas Masarik, Grzegorz Pierczynski, and Piotr Skowron. 2024. A Generalised Theory of Proportionality in Collective Decision Making. In Proceedings of the 25th ACM Conference on Economics and Computation (New Haven, CT, USA) (EC '24). Association for Computing Machinery, New York, NY, USA, 734—754. https://doi.org/10.1145/3670865.3673619

Strategic and Structural Robustness of LLM-Mediated Social Choice: Manipulation, Bias, and Control in Generative Democratic Processes

Supervisor: Maria Polukarov

Project Description

Generative social choice (Fish et al., 2024; Boehmer et al., 2025) replaces the fixed candidate sets of classical social choice with an unbounded space of LLM-generated statements, using discriminative queries (estimating agent utility for a statement) and generative queries (producing new representative statements) to build proportional slates from platforms like Polis. This literature assumes queries are answered truthfully and that the LLM is a neutral oracle, leaving open how such systems behave when participants or the model itself are adversarial. This project will bring a control/manipulation lens, previously applied only to classical fixed-candidate elections, to this infinite, query-accessible candidate space. In particular, it will: (1) formalise strategic manipulation, where agents craft inputs to bias discriminative-query estimates or steer generative queries toward self-serving statements, and characterise the computational complexity of detecting or executing such attacks; (2) treat the LLM's own training-induced biases as an implicit, unelected ``chair'' exercising candidate control over which statements can ever be generated, and developing audits/metrics for this influence; (3) propose manipulation-resistant query protocols with provable proportionality guarantees under bounded adversarial noise. The work bridges algorithmic social choice, mechanism design, and LLM evaluation.

References

Fish, S., Golz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., & Wuthrich, M. (2024). Generative social choice. In Proceedings of the 25th ACM Conference on Economics and Computation (EC 2024) (p. 985). ACM. https://doi.org/10.1145/3670865.3673547

Boehmer, N., Fish, S., & Procaccia, A. D. (2025). Generative social choice: The next generation. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). arXiv:2505.22939.

Small, C., Bjorkegren, M., Erkkila, T., Shaw, L., & Megill, C. (2021). Polis: Scaling deliberation by mapping high dimensional opinion spaces. Recerca: Revista de Pensament i Analisi, 26(2).

Jarrett, D., Pislar, M., Bakker, M. A., Tessler, M. H., Koster, R., Balaguer, J., Elie, R., Summerfield, C., & Tacchetti, A. (2023). Language agents as digital representatives in collective decision-making. In NeurIPS 2023 Foundation Models for Decision Making Workshop.

Bartholdi, J. J., Tovey, C. A., & Trick, M. A. (1992). How hard is it to control an election? Mathematical and Computer Modelling, 16(8), 27—40.
https://doi.org/10.1016/0895-7177(92)90085-Y

Hemaspaandra, E., Hemaspaandra, L. A., & Rothe, J. (2007). Anyone but him: The complexity of precluding an alternative. Artificial Intelligence, 171(5-6), 255—285.

Faliszewski, P., & Rothe, J. (2016). Control and bribery in voting. In F. Brandt, V. Conitzer, U. Endriss, J. Lang, & A. D. Procaccia (Eds.), Handbook of Computational Social Choice (pp. 146—168). Cambridge University Press.

Computational Foundations of Decentralised Markets

Supervisor: Edwin Lock

Project Description

Many markets function without a central auctioneer, relying instead on bilateral negotiation between participants who each see only part of the market. Examples include over-the-counter financial markets, supply chains, and trading mechanisms implemented on the blockchain. This project will investigate when such decentralised markets reach stable and equitable outcomes, how quickly, and at what computational cost. It will combine tools from algorithmic game theory, discrete convex analysis, and complexity theory to explain how good outcomes can arise, or provably fail to arise, when agents interact strategically in networked markets.

The overarching aim is to develop a general theory of when decentralised interaction can substitute for centralised market clearing. A first goal is to identify the structural conditions under which natural bargaining and price-adjustment dynamics converge to stable outcomes, to establish how quickly they do so, and to prove matching impossibility results where they cannot, building on recent evidence that even minimal rationality requirements can be computationally intractable. A second goal is to understand which outcomes decentralised markets actually select: how stability can be reconciled with fairness, which participants systematically lose out, and how much welfare is sacrificed relative to a central planner, using concepts such as the price of anarchy.

Underlying these questions is the role of information. The project will ask what communication is fundamentally required for decentralised markets to function, particularly where messages are costly (e.g., blockchain transaction fees) or strategically sensitive, and how robust market outcomes are to misrepresentation and adversarial behaviour. Ultimately, these results can inform the design of hybrid markets that combine the strengths of both worlds, embedding lightweight central mechanisms, such as clearing auctions, within otherwise decentralised trade.

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

Mechanism Design for Robust AI Alignment

Supervisor: Carmine Ventre

Project Description

We propose to investigate a game-theoretic 'incentive engineering' solution to AI alignment using mechanism design. A dominant strategy incentive compatible (DSIC) mechanism would align AI systems to the designer's goals (captured by a so-called social choice function) since it would incentivise AI agents to act honestly, thereby implementing the function. This ideal solution relies on the problematic assumption that AI agents are perfectly rational. The project can address this limitation in different ways.

A more theoretical approach could be to study novel solution concepts, inspired by the literature on behavioural economics. This will build on the landscape developed by Prof Ventre in his research on obviously dominant strategies. These strategies are played by agents with limited contingent reasoning skills or with access to data that is not granular enough to differentiate the payoff in each possible strategy profile. Concepts like OSP and SOSP restrict even further the class of social choice functions that are implementable vis-a-vis DSIC, by modelling agents who are only honest when it is obvious to them. NOM significantly relaxes DSIC, by assuming that agents will not manipulate the mechanism unless it is obvious to them. We will study mechanisms in the largely unexplored space between DSIC and NOM, thus guaranteeing alignment for AI agents that may find sophisticated manipulations non-obvious. We will test ideas developed at the interface of OSP and DSIC. To hope is to define a class of nested mechanisms that become more powerful as AI agents are less capable, making the results robust against advancing AI capabilities. A more empirical approach could look at empirical mechanism design for LLM agents, thereby attempting to understand how to engineer the incentives of actual AI agents operating in a multi-agent system.

Prospective applicants are encouraged to consult the publications of Prof Ventre at https://kclpure.kcl.ac.uk/portal/en/persons/carmine.ventre/publications/.

Identifying and Shifting the Frontier of Safe Multi-Agent AI

Supervisor: Carmine Ventre

Project Description

When do networks of AI agents recover safely from disruption, and when do they trigger harmful cascades? This project seeks to develop and test indicators that identify these risks, and intervention rules—changes to incentives, information flows, interaction rules, or high-risk local connections—that can prevent or contain them.

The project can attempt to study these questions from a theoretical standpoint. For example, one could study repeated network games with heterogeneous and boundedly rational agents, to model agentic AI systems that have limited context, imperfect information, finite computational resources and time, and restricted action spaces. The goal would be the characterisation of safety frontiers (in terms of, e.g., equilibria(and interventions that can shift those frontiers towards safer outcomes.

A more empirical approach could develop a reproducible multi-agent environment with adaptive agents and establish when safety is reached and which measures work reliably and where they fail.

Biologically Inspired Artificial Intelligence for Understanding and Mitigating Age-Related Cognitive Decline

Supervisor: Kathleen Steinhofel / Peter Giese

Project Description

Age-related cognitive impairment, including declines in memory, learning, and decision-making, is a growing global health challenge. While artificial neural networks (ANNs) have achieved remarkable success in modelling cognitive processes, their learning mechanisms differ fundamentally from those observed in the ageing brain. This PhD project aims to develop biologically inspired AI models based on Long-Term Potentiation (LTP), a key mechanism supporting learning and memory in younger brains, and Multi-Innervated Spines (MIS), neural structures that become more prevalent with ageing and may contribute to adaptive compensation for cognitive decline. By incorporating these mechanisms into artificial neural networks, the research will investigate how changes in synaptic plasticity influence learning speed, memory retention, adaptability, and cognitive resilience. The project seeks to provide new insights into the neural basis of age-related cognitive impairment while developing novel AI architectures capable of more robust, flexible, and energy-efficient learning. Ultimately, the work aims to bridge neuroscience and artificial intelligence, contributing both to our understanding of brain ageing and to the design of next-generation brain-inspired intelligent systems.

References

Simulating Categorical Memory Engram Formation Under Young and Ageing Brain Conditions
J Hewavitharana, PHDC Avelar, P Giese, K Steinhofel
Artificial Life Conference Proceedings 38 2026 (1), 12, 2026
Low concentrations of amyloid-beta oligomers induce synaptogenesis characteristic for mild cognitive impairment and alter the de novo proteome
K Wu, S Lee, R Martinez-Serra, L Zhang, S Lynham, KP Giese
Translational psychiatry 16 (1), 132, 2026
Aging-independent decrease of complex multi-spine boutons in hippocampal area CA1 after contextual fear conditioning
R Martinez-Serra, S Lee, I Kraev, KP Giese
Molecular brain 19 (1), 1, 2025
Brain Inspired Learning for Neural Networks
J Hewavitharana, A Anand, P Giese, C Moretti Ierardi, K Steinhofel
International Conference on Engineering Applications of Neural Networks, 59-72, 2025
Computational analysis of learning in young and ageing brains
J Hewavitharana, K Steinhofel, KP Giese, CM Ierardi, A Anand
Frontiers in Computational Neuroscience 19, 1565660, 2025

Counterfactual Causal Reasoning for Sequential Decision-Making and AI Systems

Supervisor: Nicola Paoletti

Project Description

Sequential decision-making systems such as MDPs and POMDPs underpin reinforcement learning, robotics and formal verification, but explaining why a particular decision or failure occurred remains challenging. Counterfactual causal reasoning asks what would have happened under alternative interventions, supporting tasks such as counterfactual explanation and actual-cause analysis [1].

Our group develops structural-causal-model methods for these problems. Recent work includes robust counterfactual inference for MDPs under causal-model uncertainty, extensions to data-driven uncertainty, inference for nondeterministic causal models, counterfactuals for language models, and temporal logics with counterfactual modalities [2-5].

Potential directions include:

* Causal attribution and repair for language-model agents, developing efficient algorithms for symbolic MDP abstractions and counterfactual replay. [6-9]
* Latent-level causal explanations, tracking interpretable concepts across multi-step or multi-agent interactions, with applications to explaining and mitigating misaligned AI behaviours. [10-13]
* Counterfactual analysis of learned world models, including Dreamer- and JEPA-style architectures.
* Formal verification of counterfactual properties, including numerical model checking, alternative causal semantics, continuous-state/time systems, and connections to hyperproperties [14-16].

The ideal candidate would have strong foundations in probability, machine learning, formal methods, causal inference, or reinforcement learning, and an interest in rigorous methods for understanding AI systems.

References

1. Halpern, Joseph Y. "A modification of the Halpern-Pearl definition of causality." arXiv preprint arXiv:1505.00162 (2015).
2. Lally, Jessica, et al. "Robust Counterfactual Policy Optimisation via Nondeterministic Causal Models." arXiv preprint arXiv:2608.02893 (2026).
3. Pona, Edoardo, et al. "Abstract Counterfactuals for Language Model Agents." Advances in Neural Information Processing Systems 38 (2026): 87484-87509.
4. Kazemi, Milad, Jessica Lally, and Nicola Paoletti. "Causal temporal reasoning for Markov decision processes." Research Directions: Cyber-Physical Systems 3 (2025): e3.
5. Lally, Jessica, Milad Kazemi, and Nicola Paoletti. "Robust counterfactual inference in Markov decision processes." arXiv preprint arXiv:2502.13731 (2025).
6. Beckers, Sander, Frederick Eberhardt, and Joseph Y. Halpern. "Approximate causal abstractions." Uncertainty in artificial intelligence. PMLR, 2020.
7. Lopez-Miguel, Ignacio D., et al. "ATLAS: Discovering Agent Strategies through LLM-Guided Abstraction and Automata Learning." arXiv preprint arXiv:2608.14352 (2026).
8. Bonagiri, Akash, et al. "CausalFlow: Causal Attribution and Counterfactual Repair for LLM Agent Failures." arXiv preprint arXiv:2605.25338 (2026).
9. Zhang, Shaokun, et al. "Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems." arXiv preprint arXiv:2505.00212 (2025).
10. Mueller, Aaron, et al. "The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis." Computational Linguistics 52.1 (2026): 331-378.
11. Rose, Aaron, et al. "Detecting Multi-Agent Collusion Through Multi-Agent Interpretability." arXiv preprint arXiv:2604.01151 (2026).
12. Geiger, Atticus, et al. "Causal abstractions of neural networks." Advances in Neural Information Processing Systems 34 (2021): 9574-9586.
13. Shi, Chuancheng, et al. "TraceRouter: robust safety for large foundation models via path-level intervention." arXiv preprint arXiv:2601.21900 (2026).
14. Dimitrova, Rayna, Bernd Finkbeiner, and Hazem Torfah. "Probabilistic hyperproperties of Markov decision processes." ATVA, 2020.
15. Von Kugelgen, Julius, Abdirisak Mohamed, and Sander Beckers. "Backtracking counterfactuals." Conference on Causal Learning and Reasoning. PMLR, 2023.
16. Baier, Christel, Jakob Piribauer, and Robin Ziemek. "Foundations of probability-raising causality in Markov decision processes." Logical Methods in Computer Science 20 (2024).

Algorithms for perpetual scheduling problems

Supervisor: Tomasz Radzik

Project Description

Perpetual scheduling refers to scheduling recurring events where each event must occur at a specified frequency. This task arises in various application contexts, such as determining service schedules for a collection of devices distributed across a geographical area, when each device has its own required service frequency, or scheduling medical checkups for hospital patients, when different patients require checks at different intervals.

Our goal is to design efficient algorithms for computing good (ideally optimal) perpetual schedules, while analyzing the trade-offs between computational time and the quality of the resulting schedules. There are several variants of this scheduling problem, depending on specific constraints and on how the quality of computed schedules is defined. For example, in the context of device servicing, we may aim to minimize either the maximum or the average waiting time for the next service.

This research falls within the fields of optimization, algorithm design, and computational complexity. The reference below provides further information on perpetual scheduling and the relevant computational methodologies.

References

L. Gasieniec, T. Jurdzinski, R. Klasing, C. Levcopoulos, A. Lingas, J. Min, T.Radzik, "Perpetual maintenance of machines with different urgency requirements. Journal of Computer and System Sciences, vol. 139 (2024). DOI: 10.1016/j.jcss.2023.103476.

Design and analysis of Population Protocols

Supervisor: Tomasz Radzik

Project Description

Population Protocols, introduced in the paper referenced below, are a model of distributed computing in which populations of simple, anonymous agents can perform global computational tasks. The model draws inspiration from the randomness and parallelism of natural processes and offers the potential for scalability, resilience, and adaptability.

More precisely, in the population protocol model, a large population of simple agents, indistinguishable from one another except that each may have its own initial input token, self-organises to solve computational tasks through random pairwise interactions. The states of the agents are updated according to a deterministic transition function, typically modeled as a finite-state automaton.

Over the past 20 years, extensive research on population protocols has focused predominantly on understanding the computational power of the model and on designing efficient protocols for foundational tasks in distributed computing, such as leader election and consensus. In contrast, this project aims to develop a framework for composing population protocols into higher-level programs that can provide verifiably reliable solutions to more complex distributed computing tasks.

This research falls within the field algorithm design and their mathematically-rigorous analysis.

References

[1] D. Angluin, J. Aspnes, Z. Diamadi, M.J. Fischer, and R. Peralta. Computation in networks of passively mobile finite-state sensors. Distributed Computing, 18(4):235—253, 2006.