Guillaume Corlouer

gcorlouer.github.io gcorlouer linkedin Google Scholar

Experience

AI Safety (2022-2024)

01/2024–07/2024 **Research Affiliate**
Principles of Intelligent Behavior in Biological and Social Systems (PIBBSS), London, UK
- Published at the ICML 2024 workshop on LLMs cognition on investigating information-theoretic measures for detecting deceptive outputs.
- Published a blog post on the AI alignment forum on the relevance of Bayesian statistics to predict the dynamics of stochastic gradient descent on degenerate loss landscapes.

03/2023-12/2023 **Independent researcher in AI safety**
- Published at UniReps NeurIPS workshop 2023 on discovering linear representations in transformers trained to solve mazes.
- Co-organized a workshop on AI safety and artificial life at the Alife conference 2023.
- Ranked 2nd at a mechanistic interpretability hackathon on identifying a circuit for the prediction of gendered pronouns in GPT-2 small.

05/2022–10/2022 **Independent Researcher in AI strategy** (part-time)
Contracting with Center on Long-Term Risk, London, UK
- Developed a mathematical model for optimizing philanthropic spending in AI safety.

Computational Neuroscience & Mathematics (2016-2022)

09/2018–12/2022 **Doctoral Researcher in Informatics**
Sussex Centre for Consciousness Science, School of Informatics and Engineering, University of Sussex, Brighton, UK
- Published a PhD thesis on estimating information flow between cortical regions of the human brain during visual perception.

09/2016–05/2018 **Doctoral Researcher in Pure Mathematics**
Arithmetic and Algebraic Geometry Research Group, Mathematics Laboratory, Paris-Saclay University, Orsay, France
- Conducted research in algebraic geometry & representation theory.

Fellowships

07/2024–09/2024 **Center on Long-Term Risk summer research fellowship**
- Developed a model to prioritize interventions reducing long-term catastrophic AI risk under deep uncertainty.

06/2023–09/2023 **PIBBSS summer fellowship**
- Investigated stochastic gradient descent on low dimensional loss-landscapes with broad basins of attraction.

07/2021-09/2021 **Summer Research Fellowship in AI Strategy**
Berkeley Existential Risk Initiative
- Developed a mathematical model of optimal philanthropic spending to reduce global catastrophic risks.

08/2019 **Machine Intelligence Research Institute (MIRI) Summer Fellows Program**
- Introduced to AI alignment for mathematicians, wrote a blog post on meta-ethics and AI alignment.

Publications

Proceedings

1. A.-K. Dombrowski and G. Corlouer, "An information-theoretic study of lying in LLMs," *ICML 2024 Workshop on LLMs and Cognition*, 2024.
2. M. Ivanitskiy, A. F. Spies, T. Räuker, *et al.*, "Linearly Structured World Representations in Maze-Solving Transformers," *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pp. 133–143, 2024.

PhD thesis

1. G. Corlouer, "Investigating information transfer in ECoG time series during visual perception," 2023.

Preprints and blog posts

1. G. Corlouer and N. Mace, *Degeneracies are sticky for SGD* , 2024.
2. M. I. Ivanitskiy, R. Shah, A. F. Spies, *et al.*, *A Configurable Library for Generating and Manipulating Maze Datasets*, Preprint, 2023.
3. C. Mathwin, G. Corlouer, E. Kran, F. Barez, and N. Nanda, *Identifying a circuit for gendered pronoun prediction in GPT-2 small*, 2023.
4. T. Cook and G. Corlouer, *The optimal timing of spending on AI safety work*, 2022.

Talks

1. G. Corlouer, "The role of model degeneracy on the dynamics of SGD," PIBBSS symposium, 2023.
2. G. Corlouer, "Top-down and bottom-up information flow in visually responsive neural populations," Neuromatch 2.0, 2021.

Education

2023 **PhD in Informatics**, University of Sussex, Brighton, UK

2016 **MSc in Mathematics and Applications**, Arithmetic and Geometry, Paris-Saclay University, Paris, France

2014 **MSc in Theoretical Physics**, Ecole Normale Supérieure Paris & Paris-Saclay University, Paris, France

Tech Stack

Programming — Python, LaTeX, MATLAB

Libraries — PyTorch, Pandas, NumPy, SciPy, Matplotlib

Dev Tools — Git, Github

Teaching

2017 Teaching assistant in real analysis and linear Algebra for undergraduates, Paris Saclay University

2016 Teaching assistant in linear algebra for undergraduates, Paris Saclay University

Teaching (continued)

2015 — Teaching assistant in physics for AIMS master's in mathematical sciences, African Institute of Mathematical Sciences, Mbour, Senegal

Funding

10/2023-12/2024 — Grant from Epistea to do research on AI safety as an independent researcher

03-06/2023 — Grant from Effective Ventures to work on understanding search in transformers

09/2018-12/2021 — Doctoral scholarship from the CIFAR Azrieli global scholar program for Brain, Mind, and Consciousness

2016-2018 — Doctoral scholarship from the doctoral school of mathematics Jacques Hadamard

References

Lionel Barnett
- PhD supervisor
- Research fellow at Sussex center for consciousness neuroscience, University of Sussex
- l.c.barnett@sussex.ac.uk

Anil Seth
- PhD supervisor
- Professor of Cognitive and Computational Neuroscience, University of Sussex
- Director, Sussex Centre for Consciousness Science, University of Sussex
- A.K.Seth@sussex.ac.uk

Fernando Rosas
- Colleague
- Lecturer in Computer Science and AI, School of Engineering and Informatics, University of Sussex
- F.Rosas@sussex.ac.uk

Nicolas Macé
- Collaborator
- Researcher at Center on Long-Term Risk
- n.mace@protonmail.com

Lucas Teixeira
- Research manager
- Program Lead at PIBBSS
- lucas@pibbss.ai