Guillaume Corlouer

guillaume.corlouer@gmail.com gcorlouer linkedin Google Scholar

Research Experience

AI Safety Strategy

07/2024–09/2024 **Summer Research Fellow**, *Center on Long-Term Risk, London, UK*
- Developed a model for prioritizing interventions aimed at reducing long-term catastrophic AI risk under deep uncertainty.

05/2022–10/2022 **Contracting Researcher** (part-time), *Center on Long-Term Risk, London, UK*
- Developed a model for optimizing philanthropic spending in AI safety, focusing on minimizing long-term catastrophic risk; published on the Effective Altruism Forum.

AI Safety & Interpretability

01/2024–07/2024 **Research Affiliate**
Principles of Intelligent Behavior in Biological and Social Systems (PIBBSS), London, UK
- Applied information-theoretic measures for lie detection in large language models; published in an ICML workshop.
- Investigated the influence of degenerate directions in the loss landscape on stochastic gradient descent dynamics; published on the AI Alignment Forum.

01/2023–12/2023 **Independent Researcher**
- Investigated linear representations in transformers trained to solve mazes; published at a NeurIPS workshop.
- Explored the relevance of singular learning theory for deep learning during the PIBBSS summer fellowship
- Worked on organizing a workshop on AI safety and artificial life.
- Worked on a project to identify a circuit for gendered pronoun prediction in GPT-2 small, which ranked 2nd at a mechanistic interpretability hackathon.
- Participated in the Cambridge Machine Learning bootcamp.

Neuroscience of Consciousness

09/2018–12/2022 **Doctoral Researcher in Informatics**
Sussex Centre for Consciousness Science, School of Informatics and Engineering, University of Sussex, Brighton, UK
- Estimated information flow between visual areas of the human brain to investigate conscious visual perception.
- Published a PhD thesis supervised by Anil Seth and Lionel Barnett.

Pure Mathematics

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 geometric representation theory to count principal bundles on projective curves over finite fields.
- Supervised by Olivier Schiffmann.

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 & teaching

Education

2023 — **PhD in Informatics**, University of Sussex, Brighton, UK
Thesis title: Investigating Information Transfer in ECoG Time Series During Visual Perception
Supervisors: Anil Seth and Lionel Barnett

2016 — **MSc in Mathematics and Applications**, Arithmetic and Geometry, Paris-Saclay University, Paris, France
MSc report: The Hall Algebra of Coherent Sheaves on the Projective Line
Supervisor: Olivier Schiffmann

2014 — **MSc in Theoretical Physics**, ENS Paris & Paris-Saclay University, Paris, France
MSc report: Integrable Spin Chains
Supervisor: Véronique Terras

Teaching

2016–2018 — **Teaching Assistant**
Paris-Saclay University, Orsay, France
- Taught linear algebra and real analysis to first and second year undergraduates

Education & teaching (continued)

09/2014–07/2015 **Teaching Assistant**
African Institute for Mathematical Sciences (AIMS), Mbour, Senegal
- Taught linear algebra and led tutorials in quantum mechanics, Python programming, differential geometry, and number theory
- Co-supervised two master's projects in number theory and one project in applied physics

Technical skills

Languages	Python, Matlab
Machine learning	Stochastic gradient descent
Statistics	Non-parametric hypothesis testing, Granger causality, Singular learning theory, State-space models, Vector autoregressive modeling
Information theory	Transfer entropy, Information decomposition, measures of Emergence

Funding

10/2023-01/2024	Grant from Rory Greig to do research on AI safety as an independent researcher, £10,000
03-06/2023	Grant from Effective Ventures to work on understanding search in transformers, £5,000
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