LENNIE WELLS

Contact: ww347@cam.ac.uk GitHub: https://github.com/W-L-W
Google Scholar LinkedIn

PROFILE

Final-year Cambridge PhD in Statistics/ML and core member of the Cambridge AI Safety research ecosystem.

Published work on RLHF, self-supervised learning, and high-dimensional statistics. Currently pursuing empirical research on LLM self-recognition, AI debate, and low-stakes control. High integrity individual seeking bridge to full time AI safety research role and maximal impact.

EDUCATION

· **PhD in Statistics and Machine Learning** Oct 2021 – Present
University of Cambridge (Cambridge, UK)

Supervised by Sergio Bacallado. Canonical correlation analysis from perspectives on high-dimensional statistics and self-supervised learning; foundations of Bayesian inference and neural processes; reinforcement learning for language models.

· **MMath specialising in probability and statistics** Oct 2020 – Jun 2021
University of Cambridge (Cambridge, UK). Distinction. Rank 10 / 272. Mark 94%.

· **BA in Mathematics** Oct 2017 – Jun 2020
University of Cambridge (Cambridge, UK). High first class (and top mark in college) each year of examination.

AI SAFETY EXPERIENCE

· **Co-worker**. *Meridian Office* (Cambridge, UK) Oct 2024 – Present
Daily collaboration on AI safety research and regular contributions to group discussions and lightning talks.

· **Co-organiser and Facilitator for Research Accelerator Week**. *Meridian* (Cambridge, UK) Apr 2025
Facilitated and co-organised Meridian Visiting Research Week for 20 researchers forming OpenPhil grant teams. Ran workshops on project ideation and goal-oriented research. Secured special agreement with OpenPhil for priority EOI review mid-week.

· **Co-organiser & Lecturer for AI Safety Course**. *Meridian/C2D3* (Cambridge, UK) Jan – Mar 2025
Co-organised and lectured for 16-lecture series on 'Language Models and Intelligent Agentic Systems'. Recordings average 400+ views per lecture.

· **MARS 3.0 Mentor**. *MARS/Geodesic Research* (Cambridge, UK) Jun 2025 – Present
Mentoring project on LLM Self-Recognition as part of Geodesic Research mega-stream (see below).

· **External Collaborator**. *Mary Phuong's MATS 8.0 Stream* (Cambridge & London, UK) Jun – Aug 2025
Research on black-box detection methods for low-stakes AI control (see below).

ONGOING WORK AND RESEARCH INTERESTS

· **Black-box detection for low-stakes control** Jun-Aug 2025
Constructing an adversarial game for detection of strategic underperformance, with primary application to research sabotage threat models. Main contributions around game design and consistency-based blue team strategies.

· **Train-time oversight detection** Jun 2025 - Present
Constructing model organisms to investigate whether LLMs may be able to infer the extent of meaningful oversight via mechanisms related to out-of-context learning.

· **Self-recognition in LLMs** Jun 2025 - Present

Investigating different operationalisations of LLM self-recognition, with the aim of understanding which notions of self are relevant to the different applications in alignment and control. Submitted AISI Challenge Fund application.

· **AI Debate** Aug 2025 - Present
De-risking a planned 12 month empirical project on AI debate, for AISI Alignment Project application.

PUBLICATIONS

· **L Wells**, E Young, J Brown, S Bacallado. KL-Regularised Q-Learning: A Token-level Action-Value perspective on Online RLHF. 2nd Workshop on Models of Human Feedback for AI Alignment, ICML (2025).

· G Flamich, **L Wells**. Some Notes on the Sample Complexity of Approximate Channel Simulation. Spotlighted in *Learn to Compress Workshop*, *ISIT* (2024).

· **L Wells**, K Thurimella, S Bacallado. Regularised Canonical Correlation Analysis: graphical lasso, biplots and beyond. *arXiv preprint arXiv:2403.02979* (2024).

· J Chapman, AL Aguila, **L Wells**. Unconstrained Stochastic CCA: Unifying Multiview and Self-Supervised Learning. *ICLR* (2024).

TECHNICAL SKILLS

· **Safety Research**: Designing and implementing evals and control evals (Phuong project). Finetuning via APIs. RLHF with trl (see KL-Regularised Q-Learning). Broad knowledge and deep understanding of AI safety literature.

· **Python**: numpy, matplotlib, scikit-learn, pandas, pytorch, huggingface, inspect.

· **Workflow**: VSCode/Cursor, Vim, Git, bash, GNU/Linux, LaTeX, Weights & Biases.

· **Languages**: English (native), French (B1/2), German (B1/2).

ADDITIONAL EXPERIENCE

· **Quant Research Intern**. *G-Research* (London, UK) Jun 2024 – Aug 2024
Constructed cross-sectional equities signal using various data feeds and ML techniques.

· **Quant Research Intern**. *Capula Investment Management* (London, UK) Jun 2023 – Aug 2023
Developed end-to-end futures strategy utilizing ChatGPT for sentiment analysis of central bank speeches.

· **Trading Intern**. *Susquehanna International Group* (Dublin, Ireland) Jun 2021 – Aug 2021
Completed intensive training in options theory, mock trading, and poker strategy. Conducted research project on SX5E index options.

· **Statistics Consultant**. *University of Cambridge* (Cambridge, UK) Oct 2021 – Present
Provide statistical consulting for interdisciplinary projects in neuroscience, linguistics, and engineering.

· **Teaching Assistant/Supervisor**. *University of Cambridge* (Cambridge, UK) Oct 2021 – Present
- Analysis and Topology (2nd year)
- Mathematics of Machine Learning (3rd year)
- Principles of Statistics (3rd year)
- Teaching Assistant for Bayesian Modelling and Computation

AWARDS

· **Larmor Award** Jun 2021
One of only seven undergraduates to receive prestigious college prize for "intellectual qualifications, moral conduct and practical activities".

· **College Prize and Baylis Scholarship for Mathematics** 2018, 2019, 2020, 2021
For high first class and top mark in college in all years of undergraduate and master's mathematics exams.

REFERENCES

Available upon request.