John Burden

(+44)7713811215 ⋄ jjb205@cam.ac.uk ⋄ johnburden.co.uk

EMPLOYMENT

Leverhulme Centre for the Future of Intelligence, University of Cambridge

Programme Co-director: Kinds of Intelligence — January 2024 - Present
Senior Research Associate — June 2023 - Present
Research Associate — January 2022 - June 2023

- P.I on Solid Foundations project, where I am investigating risks that arise from Foundation Models, their interactions, and ubiquity in society.
- Post-Doctoral Research Associate on the RECOG-AI project. This aims to develop novel methods to assess capabilities and the generality of AI systems using approaches inspired by Cognitive Science.

Centre for the Study of Existential Risk, University of Cambridge

Senior Research Associate — June 2023 - Present
Research Associate — July 2020 - June 2020

- Post-Doctoral Research Associate on the FLI-funded project: Paradigms of Artificial General Intelligence and Their Associated Risks.
- Contributor to the AI: Futures and Responsibilities team and Kinds of Intelligence group

Sidney Sussex College, University of Cambridge

College Research Associate — January 2024 - Present

- Nominated for college affiliation, providing supervisions for Computer Science students.

EDUCATION

University of York — September 2017 - August 2021
Doctor of Philosophy in Computer Science

University of Oxford — October 2013 - July 2017
Master's of Computer Science — First Class

RESEARCH

Solid Foundations — 2024 - Present

- Extending AI evaluation techniques to account for populations of AI models.
- Developing a framework for analysing interactions between groups of AI systems, and between AI systems and users.
- Two academic publications in preparation.

RECOG-AI — 2021 - Present

- 'Robust Evaluation of Cognitive Capabilities and Generality in Artificial Intelligence'
- Developing framework for AI evaluation based on cognitively-defined capabilities drawing on Psychometrics and Psychology.
- Six academic publications, one of which is under review.

Paradigms of Artificial General Intelligence and Their Associated Risks 2020 - 2024

- Investigating the relationships between system capability, generality and risk within AI systems.
- Devised methods to make existing reinforcement learning algorithms safer using the agent's perceived self-confidence.
- Developed framework for robust evaluation of multi-modal foundation models with a focus on accounting for cognitive efforts by users as well as ensuring safety.
- Four academic publications.

Safe Reinforcement Learning for Sepsis Treatment 2020

- Adapted a deep reinforcement learning algorithm with policy-constraints to improve the safety of AI-proposed sepsis treatments.
- Resulted in one conference publication which was subsequently extended into a journal publication.

Abstraction Within Reinforcement Learning 2017 - 2021

- PhD Project titled "Automating abstraction for potential-based reward shaping"
- Three academic publications on improving the convergence speed of Reinforcement Learning using agents that automatically generate abstractions from interaction with their current environment, using these to improve the efficiency of learning.

ACADEMIC SERVICE

Peer Review

- PC Member: ECAI 2023, ECML/PKDD 2023, AISafety 2023, SafeAI2023, AIES22, AISafety 2022. AAMAS 2022 (Innovative Applications track), SafeAI2022, ECML/PKDD 2021, AAMAS 2021, ALA 2021, ECAI 2020, AAAI-2020 (student poster programme)
- Sub-reviewer: IJCAI 2019, AAMAS 2019

Cambridge Existential Risk Initiative Mentor

- Mentored a CERI fellowship investigating improving cooperation in Iterated Prisoner's Dilemmas
- Lead to an accepted workshop paper for the student at SafeAI2022

Co-organiser of AI Evaluation Beyond Metrics (EBeM) Workshop

- EBeM accepted papers and hosted talks focused around improving AI evaluation with cognitively inspired approaches.
- Published proceedings of the accepted papers.

Co-organiser of Predictable AI Launch Event

- Predictable AI launched a new initiative to form a community of AI researchers and policy-makers to explore questions surrounding approaches to making AI systems more predictable.

GRANTS AND ACCOLADES

Long-Term Future Fund

- Awarded £209,501 for Solid Foundations project.

FLI Worldbuilding Competition

- Led the CSER/CFI submission to FLI's aspirational AI worldbuilding contest where our submission was awarded the 2nd place prize.

PUBLICATION HIGHLIGHTS

- Inferring Capabilities from Task Performance with Bayesian Triangulation. **John Burden**, Konstantinos Voudouris, Ryan Burnel, et al. arXiv preprint arXiv:2309.11975, Under Review JMLR. 2023
- Predictable Artificial Intelligence. Lexin Zhou, Pablo A. Moreno-Casares, Fernando Martínez-Plumed, **John Burden**, et al. arXiv preprint 2310.06167 2023
- Harms from Increasingly Agentic Algorithmic Systems. Alan Chan, Rebecca Salganik, ... **John Burden**, et al. 2023. Harms from Increasingly Agentic Algorithmic Systems. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 651–666. https://doi.org/10.1145/3593013.3594033
- Rethink Reporting of Evaluation Results in AI. Ryan Burnell , Wout Schellaert, **John Burden**, et al. Science 380,136-138(2023).DOI:10.1126/science.adf6369 2023
- Gruetzemacher, Ross, Alan Chan, Kevin Frazier, Christy Manning, Stepán Los, James Fox, Jos'e Hern'andez-Orallo, **John Burden**, Matija Franklin, Clíodhna Ní Ghuidhir, Mark Bailey, Daniel Eth, Toby D. Pilditch and Kyle A. Kilian. "An International Consortium for Evaluations of Societal-Scale Risks from Advanced AI." ArXiv abs/2310.14455 (2023).
- Not a Number: Identifying Instance Features for Capability-Oriented Evaluation. Ryan Burnell, **John Burden**, Danaja Rutar, Lucy Cheke, José Hernández-Orallo, Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022
- How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild. P A M Casares, Bao S. Loe, **John Burden**, Seán Ó hÉigeartaigh, José Hernández-Orallo, Proceedings of the AAAI Conference on Artificial Intelligence, 2022
- Evaluating Object Permanence in Embodied Agents using the Animal-AI Environment. Konstantinos Voudouris, Niall Donnelly, Danaja Rutar, Ryan Burnell, **John Burden**, José Hernández-Orallo and Lucy G. Cheke. EBeM'22: Workshop on AI Evaluation Beyond Metrics, July 25, 2022, Vienna, Austria
- Negative Side Effects and AI Agent Indicators: Experiments in SafeLife. **John Burden**, José Hernández-Orallo, Seán Ó hÉigeartaigh, SafeAI@AAAI 2021
- Latent Property State Abstraction For Reinforcement Learning. **John Burden**, Sajjad Kamali Siahroudi, Daniel Kudenko. In Workshop on Adaptive Learning Agents (ALA) at AAMAS 2021
- Safety-driven design of machine learning for sepsis treatment. Yan Jia, Tom Lawton, **John Burden**, John McDermid, Ibrahim Habli, Journal of Biomedical Informatics, Volume 117, 2021, 103762, ISSN 1532-0464,
- Automating Abstraction for Potential-based Reward Shaping (Thesis). **John Burden**. Whiterose e-thesis repository. 2020.
- Exploring AI Safety in Degrees: Generality, Capability and Control. **John Burden** and José Hernández-Orallo, SafeAI@AAAI2020
- Safe Reinforcement Learning for Sepsis Treatment. Yan Jia. **John Burden**, Tom Lawton, Ibrahim Habli (2020) In: 8th IEEE International Conference on Healthcare Informatics
- Uniform State Abstraction For Reinforcement Learning. **John Burden** and Daniel Kudenko. In proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)
- Using Uniform State Abstractions For Reward Shaping With Reinforcement Learning. **John Burden** and Daniel Kudenko. In Workshop on Adaptive Learning Agents (ALA) at the Federated AI Meeting. 2018

TEACHING

Evaluation of AI Systems: Capabilities, Safety and Generality Lent Term 2024

- Elective course for LCFI's MPhil degree called Ethics of AI, Data and Algorithms.
- I developed the syllabus and gave all lectures for the course.
- Highly rated course on the MPhil with great student feedback. This course is planned to become a permanent fixture on the MPhil.

Technical Foundations of AI Michaelmas Term 2024

- Development and (upcoming) teaching of one of two core courses for LCFI's MPhil degree providing an introduction to the technical foundations of AI.

Computer Science Supervisions Easter Term 2024

- Provided supervisions in Artificial Intelligence for Cambridge students from Sidney Sussex, Lucy Cavendish, Newnham, and St. Catherines colleges.

TECHNICAL SKILLS

Machine Learning	Deep Learning with Tensorflow and Keras, Reinforcement Learning, Language Models, Computer Vision
Data Analysis	Python, NumPy, Pandas, PyMC, PyPlot, OpenCV, D3.js
Software Development	Java, Scala, Haskell
Miscellaneous	LaTeX, Slurm, Bash, Vim

PERSONAL PROFILE

I am a meticulously organised and highly analytical individual. I'm proactive in my work as well as an effective and efficient communicator. In my spare time I enjoy playing board games and I also practice Historical European Martial Arts. I am an avid reader with a keen interest in history, philosophy, psychology, and speculative fiction, which has contributed to an aspirational vision for humanity and what it can achieve in the near future.