John Burden

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

EMPLOYMENT

Centre for the Future of Intelligence, University of Cambridge January 2022 - Present
Research Associate

- 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 July 2020 - Present
Research Associate

- 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

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

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.
- Two academic publications with more awaiting submission

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

- 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, one currently under review as well as one more awaiting submission and available upon request.

Safe Reinforcement Learning for Sepsis Treatment 2020

- A side project conducted in tandem with the primary PhD research.
- 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 and then use these to improve the efficiency of learning.

ACADEMIC SERVICE

Peer Review

- PC Member: 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

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.

ACCOLADES

FLI Worldbuilding Competition

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

PUBLICATIONS

- 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
- Oases of Cooperation: An Empirical Evaluation of Reinforcement Learning in the Iterated Prisoner’s Dilemma. P Barnett, J Burden - SafeAI@AAAI, 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
- Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, ..., John Burden et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615, 2022.

- 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

TECHNICAL SKILLS

Machine Learning	Deep Learning with Tensorflow and Keras, Reinforcement Learning, Language Models, Computer Vision
Data Analysis	Python, NumPy, Pandas, 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, and speculative fiction, which has contributed to a very long-term outlook on the future and potential of humanity.