Tom McGrath

thomas.m.mcgrath@gmail.com

Experience

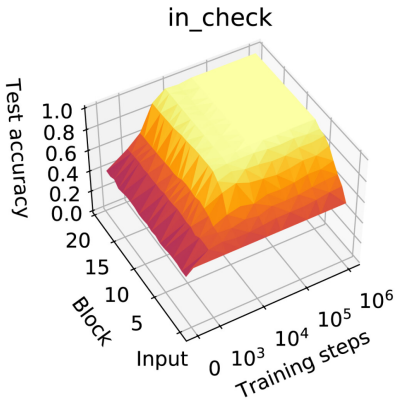
Google DeepMind — **August 2019 - August 2023**
Senior Research Scientist — *London, UK*

- Interpretability research
  - Co-founded DeepMind's interpretability team, presented research to senior leadership.
  - Led research project investigating concept acquisition trajectories in the AlphaZero deep RL agent.
  - Studied the effect of targeted ablations on neural network computations, discovered unexpected self-repair properties.
  - Built automated circuit discovery and self-interpretability tools for large language models.
- Large language model research
  - Built tooling for semantic search on LLM training data to study the effects of training data & prompt engineering.
  - Contributed to the Sparrow project by analysing human preference data and investigating model-written evaluations.
- Evaluation of generalist deep reinforcement learning agents
  - Ran agent evaluation for a large-scale general intelligence project, redesigned evaluation metrics for statistical robustness and presented across the company.
  - Organised general intelligence reading group.

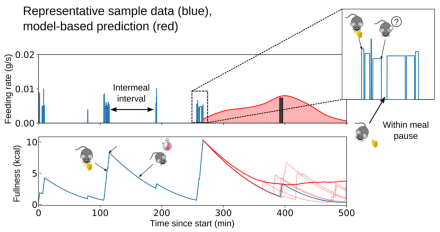
Google DeepMind — **July 2018 - December 2018**
Research Scientist Intern — *London, UK*

Future of Humanity Institute, Oxford University — **January 2018 - June 2018**
Research Scientist Intern — *Oxford, UK*

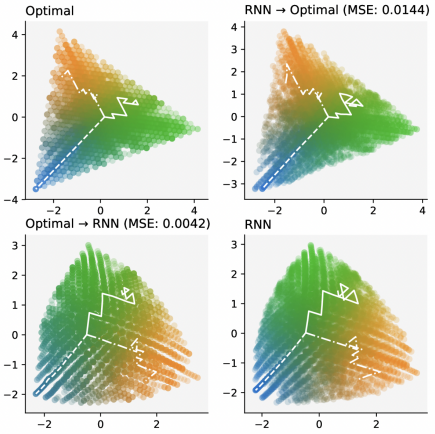
Selected Publications



Acquisition of chess knowledge in AlphaZero, *PNAS 2022.* First study demonstrating that a superhuman AI system has human comprehensible representations.



The homeostatic dynamics of feeding behaviour identify novel mechanisms of anorectic agents, *PLOS Biology 2019.* Derived a new model class for high-resolution behavioural analysis, used to uncover behavioural signatures of different drugs.



Meta-trained agents implement Bayes-optimal agents, *NeurIPS, 2020.* Analysed internals and behaviour of meta-trained neural networks.

Education

Imperial College London — **September 2014 - August 2019**
PhD in Mathematics, Doris Chen Merit Award winner — *London, UK*

- Research interests: applied Bayesian modelling; stochastic thermodynamics of computation.

Imperial College London — **September 2013 - August 2014**
MRes in Mathematics: Distinction — *London, UK*

- Relevant modules: Computational Stochastic Processes, C Programming

Warwick University — **September 2006 - July 2010**
Masters in Mathematics & Physics: 1st class honours — *London, UK*