Joar Skalse

joar.mvs@gmail.com + 44 7392957774

EDUCATION

DPhil – Computer Science
Oxford University, 2020-present
Thesis Topic: Safe Reinforcement Learning
Affiliations: FHI DPhil Scholar, member of OXCAV, affiliate of WhiRL
Supervisor: Alessandro Abate

MCompPhil – Computer Science and Philosophy
Oxford University, 2019-2020
Grade: First Class (rank 1 in the year)
Dissertation: Lexicographic Multi-Objective Reinforcement Learning. In this project I introduced reinforcement learning algorithms that accept multiple reward signals, and learn a policy that maximise the rewards lexicographically.
Extended Essay: What, if Any, Are the Advantages of Connectionism Over the Classical Computational-Representational Theory of Mind?
Awards: The Hoare Prize in CS and Philosophy for best overall performance.

BA – Computer Science and Philosophy
Oxford University, 2016-2019
Grade: First Class (rank 3 in the year)
Projects: A computer vision group project in collaboration with Microsoft. My contribution involved some data augmentation and transfer learning.

Machine learning courses taken:
Machine Learning, Computational Learning Theory, Theories of Deep Learning.

RESEARCH EXPERIENCE

The 2018 Gran Canaria AI safety camp
Paper: Reinforcement Learning in Newcomblike Environments. (under review)

The 2018 MIRI Summer Fellows Programme
Paper: Risks from Learned Optimization in Advanced Machine Learning Systems.

The 2018 Hertford Research Studentship
I investigated methods for improving the performance of program synthesis systems based on inductive logic programming, supervised by Dr. Andrew Cropper. The investigated methods did not produce significant performance gains.

Research with the Louis Group
Paper: Neural networks are *a priori* biased towards Boolean functions with low entropy.
Paper: Is SGD a Bayesian sampler? Well, almost. (under review)

My MCompPhil Dissertation
Paper: Lexicographic Multi-Objective Reinforcement Learning (under review)

Other Research
Paper: Safety Properties of Inductive Logic Programming (AAAI 2021)
Paper: A General Counterexample to Any Decision Theory and Some Responses

CODING

Implemented models (non-exhaustive): DQN, RAINBOW, RCPO, AproPO, risk-constrained RL, Neural Style Transfer, Deep Dream, Fast Gradient Sign Manipulation, DCGAN, Generative Adversarial Active Learning.

Other programming (non-exhaustive): Inductive Logic Programming in Prolog, machine learning with Judea Pearl-style causal models, a modal logic SAT-solver, large numbers of small machine learning experiments, considerable amounts of programming for coursework (including many classical AI algorithms).

COURSES

Graduate:

Computer Science: Category Theory, Computational Learning Theory, Automata Theory, Theories of Deep Learning

Philosophy: Philosophy of Cognitive Science, Decision Theory.

Undergraduate:

Programming: Functional Programming, Imperative Programming I & II.

Computer Science: Machine Learning, Artificial Intelligence, Introduction to Algorithms, Advanced Algorithms and Data Structures, Models of Computation, Concurrency, Complexity Theory.

Computational Logic: Knowledge Representation and Reasoning, Computer-Aided Formal Verification, Logic and Proof.

Mathematics: Discrete Maths, Set Theory, Information Theory, Probability.

Philosophy: Introduction to Philosophy, Turing on Computability and Intelligence, Knowledge and Reality, Philosophy of Science.

Philosophical Logic: Introduction to Logic, Elements of Deductive Logic, Philosophical Logic.