Lee Sharkey BA Hons. (Cantab.), MSc.

leedsharkey@gmail.com

Overview

2020-2021 **Researching methods for the interpretation of artificial agents**
Funded by grant from Centre for Effective Altruism (Long Term Future Fund).

2018-2021 **MSc Neural Systems and Computation, Institute of Neuroinformatics, UZH & ETHZ.**
Thesis project in the Mante Lab; TA for ETHZ course *Learning in deep ANNs and BNNs*

2017-2018 **MSc in Data Analytics, University of Glasgow.** Class: Distinction

Mid-2017 **Commenced career transition to AI and computational neuroscience.**

2016-2017 **Public health consultant, World Health Organisation, Geneva.**

2011-2014 **BA Hons. (Cantab). Pembroke College, University of Cambridge.**
Part II Neuroscience. Part I Preclinical Medical and Veterinary Sciences.
Class 2:1 with first class mark in research project.

Education

2018-2021 **Institute of Neuroinformatics, UZH and ETHZ, NSC MSc Candidate**
Nine-month full time research project (see project section below).
Deep Reinforcement Learning Seminar; Introduction to Neuroinformatics; Deep Learning; Computational Vision; Neural Systems; Systems Neuroscience; Basics of Instrumentation; Readings in Neuroinformatics; Models of Computation; Journal club.

2017-2018 **University of Glasgow, MSc in Data Analytics:** Class - Distinction
Artificial Intelligence; Probability; Bayesian Statistics; Regression Models; Statistical Inference; Big Data Analytics; Data Analysis; Generalised Linear Models; Machine Learning; Stochastic Processes; Dissertation (see projects below).

2014-2015 **Barts and the London School of Medicine and Dentistry, Clinical Medical Studies**
Discontinued study in order to take consultancy with World Health Organisation.

2011-2014 **Pembroke College, University of Cambridge**

2013-2014 Part II Neuroscience (Natural Sciences Tripos): Class - 2:1
Control of Action; Sensory Transduction; Neural Degeneration and Regeneration; Central Mechanisms of Sensation and Behaviour.
Part II Research Project (Awarded First class): *RNAi Knockdown of Phosphatidylinositol 4-Kinase IIIα in* Drosophila Melanogaster *and its effect on phototransduction*.

2011-2013 Part I (Medical and Veterinary Sciences Tripos):
Physiology, Biochemistry, Anatomy, Statistics for Medicine, Social Context of Health and Illness. Pathology, Pharmacology, Neurobiology and Medical Psychology, Human reproduction. Electives: Experimental Psychology; Sensorimotor Neurobiology.

2004-2011 **Methodist College Belfast**

2011-2010 A-levels: Biology (A*), Chemistry (A*), Physics (A*). AS-level: French (B), Maths (A [max])

2009 GCSEs: 8A*, 3A; Third in Northern Ireland in Business Studies.

Computational Skills

Github

https://github.com/leesharkey/

Languages

Python (Strong), R (Intermediate).

Skills

Deep learning and Machine learning: Have implemented projects in Pytorch, Tensorflow, and Jax deep learning libraries. *Reinforcement learning:* Have trained and analysed agents. *Distributed computing:* Have used distributed processes on Euler high-performace computing cluster). *Statistical data analysis:* Through various projects and education.

Projects

On the computational function of context-dependent gamma synchrony.

Linked here. Nine month full time MSc thesis. Written in Python using Pytorch. I conceived and worked on the project entirely independently.

Abstract: Cortical computation is carried out by transiently activated sets of neurons whose activation is structured temporally by oscillations. A classic and surprising result in cortical neuroscience revealed that gamma oscillations of nearby neurons in primary visual cortex with the same orientation preference synchronise when stimulated with colinear stimuli. Qualitative explanations suggest that synchrony implements contour integration, since edges are frequently colinear in natural scenes, and hence reflects correlation. But so far no model has bridged the gap between correlation and synchronous neural dynamics. Recently, it has been suggested that cortical gamma oscillations may implement efficient probabilistic inference through dynamics that resemble Hamiltonian Markov Chain Monte Carlo. This thesis asks whether this probabilistic account of cortical oscillations can extend to context-dependent gamma synchrony. If such synchrony were to emerge in a model (trained to represent natural scenes) that performs Hamiltonian inference over its dependent latent variables, as is hypothesised in cortex, then it would provide the correlational account of context-dependent gamma synchrony with a quantitative model of its emergence. Here we train such a model. However, we find no synchrony in such a model, suggesting that context-dependent synchrony results from factors beyond mere correlation using Hamiltonian inference.

An investigation of the neural dynamics of meta-reinforcement learners in visual decision making tasks and comparison with primate prefrontal cortical neural dynamics.

Initial topic of MSc thesis in the Mante lab at INI. Put on pause because a group at DeepMind were conducting a nearly identical investigation and I wanted time to reflect on

interesting tangential directions. Four months full time; written in python using JAX deep learning library. Preliminary results were presented to members of the DeepMind Neuroscience and Safety teams. The project investigated the neural dynamics of meta-reinforcement learners (Duan et al., 2016; Wang et al., 2016) by 'breaking open the black box of RNNs' (Barak and Sussillo, 2013) in order to better characterise the inner RL algorithm and find candidate hypotheses to explain PFC recordings from primates trained to perform a visual decision making task. In follow-on work, I plan to characterise the inner RL algorithm by linking neural dynamics to components of human-designed RL algorithms and to study learned algorithms for recurrent implementations of causal reasoning, planning, and model-based RL. My aim is to shed light on mesa-optimizers in agents (specifically on the structure and objectives of their learned inner algorithms) to begin to map the space for safety-monitoring efforts.

Investigating the emergence of grid-cell representations (GCRs) in artificial agents.

For personal curiosity; written in Python using Tensorflow. Adapting code from Banino et al. (2018), who found that, for unknown reasons, only with dropout did an artificial neural network learn GCRs. I explored possible reasons (in particular, sparsity and/or minimisation of representations' dependence on single directions) and alternative ways to yield GCRs. My results indicate that GCRs result from minimisation of representations' dependence on single directions by the introduction of any noise (not just binary dropout).

Cortically-inspired hierarchical predictive unsupervised learning with disentangled representations.

For ETH Deep Learning course; Written in Python using Tensorflow. Investigated an LSTM-based model of the cortical visual hierarchy for video frame prediction using Tensorflow.

Kaggle-based project For Glasgow MSc course 'Big Data Analytics'; python/keras. ConvNet-based classifier of time series data of stars with exoplanets. Download Jupyter notebook on Github.

Glasgow MSc Dissertation: *Deconvolving mutational signatures from generated cancer genome copy-number mutation data* - application of non-negative matrix factorisation and hierarchical dirichlet processes to generated cancer genome data to evaluate the algorithms' capacity to deconvolve cancer mutational signatures in R and Python. PDF available on Github.

Employment

2019-present **Independent researcher investigating principles, challenges, and methods of artificial agent analysis for the purpose of continuous monitoring and control**

Funded by a research grant from the Centre for Effective Altruism Long Term Future fund. The overall research objectives are (1) Conduct a safety-oriented literature review of artificial and biological agent analysis in order to lay out the tools that are currently available and explicate the lessons learned so far. (2) Identify principles and challenges of

monitoring and interpreting the internal processes of artificial agents for the purposes of control, paying particular attention to the challenges posed by highly capable AI agents. (3) Develop new methods for continuously or predictively monitoring and analysing artificial agents for the purposes of control.
Current research primarily works toward objective (2) and focuses on comprehensively understanding a simple meta-reinforcement learning agent.
Grant amount: $44,668. Grant report here.

2019-present **Teaching assistant, ETHZ course "Learning in Deep Artificial and Biological Neural Networks", by Prof. B. Grewe.**
Contracted to write, mark, and present an exercise for a course component "Deep learning with spikes". Written in Pytorch.

2017-2018 **Contractor, Management of Noncommunicable Diseases unit, *World Health Organization***
Wrote the *WHO Guidelines for the Medical Management of Cancer Pain in Adults and Adolescents*.

2016-2017 **Consultant, Management of Noncommunicable Diseases unit, *WHO, Geneva***
Coordinated the development of *WHO Guidelines for the Medical Management of Cancer Pain in Adults and Adolescents*. Provided general support to WHO's palliative care work, including article 'The Preventable Pain Pandemic'.
During this role, in an unofficial capacity, I highlighted this cause to the EA community by writing 'Increasing access to Pain Relief in Developing Countries – An EA Perspective', which was, at its peak, the fifth most upvoted post of all time on the forum and is routinely highlighted as an important peice of EA writing. The post helped galvanise further EA work in this area.

2015-2016 **Consultant, Human Reproduction Team, *WHO, Geneva***
Supported the development of clinical management guidelines for Chlamydia, Syphilis, Gonorrhoea, and Herpes; an epidemiological model of STIs; and the WHO STI global strategy; STI surveillance data management and reporting. Supported Ebola Virus Persistence Study by: conducting a literature review; an international meeting of stakeholders; and developing study protocol amendments for submission to WHO Ethical Review Committee.

2015 **Research Assistant, Ebola Virus Persistence Study, *WHO, Freetown, Sierra Leone***
Set up a study in a resource limited setting in a major public health emergency working for the organization leading the global response.

2014 **Intern, Human Reproduction Team, *WHO, Geneva***
Automated global STI data cleaning using VBA programming and produced *'A tool for the strengthening of STI surveillance at the country level'* and the first draft of a validation tool and checklist for *'Human Rights, gender equality, and engagement of civil society in the Elimination of Mother-to-Child Transmission of HIV and Syphilis'*.

Publications

2017 **National Palliative Care Capacities Around the World: Results from the World Health Organization Noncommunicable Disease Country Capacity Survey**. Sharkey L et al. (2017) Palliat Med. PubMed: 28677988 DOI: 10.1177/0269216317716060

Other selected writing

2019 **Systems Neuroscience for AI: An introductory guide to the literature**

A series of twelve blog posts on my github website that summarises a selection of my readings from late 2018 and early 2019. The guide aims to give AI researchers a skeleton-frame introduction to systems neuroscience and contextualises some recent AI research with respect to those systems. Titles: 1) Introduction; 2) Motivation: why should we pay attention to the brain for AI?; 3) Overviews: What, broadly speaking, is the brain computing? 4) The Cerebral Cortex: A Very Tangled Web; 5) The Thalamus: More than just Central Station; 6) The Hippocampus: A Less Tangled Web; 7) Cortico-hippocampal interactions; 8) Reinforcement Learning with the Basal Ganglia and prefrontal cortex; **9) (Recommended sample) The Forebrain: Or, How I Learned Concepts in the Cortico-thalamo-basal ganglia-hippocampal system**; 10) The 'Little Brain', often forgotten: the Cerebellum; 11) The Cerebello-basal ganglia-thalamo-cortical system; 12) Conclusions

Presentations

2020 **Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks, Mastroguiseppe and Ostojic (2017).** Institute of Neuroinformatics Journal Club paper.

2020 **Model Based Reinforcement Learning: The State of the Art; the Blurred Edges of MBRL; MBRL in nature.** Seminar in Deep Reinforcement Learning.

2016 **WHO's Approach to Strengthening Palliative Care & the Importance of Spirituality**

November 2016, Eleventh International Conference in Spiritual Care, Catholic University of Korea, Seoul, a WHO Collaborating Centre for Palliative Care

2016 **WHO's Approach to Strengthening Palliative Care**

October 2016, Société Internationale d'Oncologie Pédiatrique (SIOP) 48th Annual congress, Dublin

Extracurriculars

2012-present **Involvement in the Effective Altruism movement**

Former committee member and speaker for Effective Altruism Geneva (EAGe). Signatory of the Giving What We Can Pledge. Helped promote 'lack of access to pain medication' to a mainstream cause area in Effective Altruism.

2015 **Certificate of Merit, Joy and David Long Essay Prize, titled "Should Politics Interfere with Medicine?"**

2013 **Runner-up, Dame Ivy Compton-Burnett Prize for Creative Writing (Pembroke College, Cambridge)** for "*Sonata for Truly Unaccompanied Achilles*", a Gödel Escher Bach fan-fiction.

Model United Nations Delegations:

2011 Lady Eleanor Holles MUN Conference, Environment committee

2010 Methodist College Belfast MUN Conference, Health committee, Highly Commended

Sport and Music

2009-2014 Interprovincial track & field and cross-country running. Intercollegiate track & field, skiing, rugby.

2009-2014 Pembroke College Chapel Choir; St. Peter's Cathedral Schola Cantorum; Methodist College Belfast's Chapel Choir.

2014 Proud member of and masseur for Cambridge Massage Society