Academic Curriculum Vitae for

Hamish Haggerty

website: hamish-haggerty.github.io

Academic Qualifications

Masters in Statistics 2023
University of New South Wales, Australia

Diploma of Arts 2015
University of Queensland, Australia

Extended Major	Mathematics

Bachelor of Arts 2012
University of Queensland, Australia

Extended Major	Psychology
Minor	Logic and Philosophy of Science
Additional courses	5 Mathematics electives

Academic Highlights

University of Queensland
GPA 6.75 / 7 for Mathematics courses at 3rd and 4th year level
Four Deans Commendation awards for high academic achievement

Australian National University
Summer Research Scholarship 2016

Project	Algebraic Topology and Homotopy Theory
Supervisor	Dr Vigleik Angeltveit

University of NSW
Neural Networks and Deep Learning: 88/100
Advanced Machine Learning: 88/100
Bayesian Inference and Computation: 84/100
WAM: 81.700

Thesis: 2023

Project	*Self-Supervised Framework to Address Limited Data for Cancer Diagnosis*
Supervisor	Dr Rohitash Chandra
Grade	91/100
Github	https://github.com/hamish-haggerty/cancer-proj

Publications:
Haggerty, H., & Chandra, R. (2023). *Self-Supervised Learning for Skin Cancer Diagnosis with Limited Training Data.* Under review at Computer Methods and Programs in Biomedicine.

In my thesis work at UNSW I studied the problem of cancer image classification in the low labelled data regime. The standard technique in this regime is transfer learning, with the base network usually pretrained on ImageNet through supervised learning. An alternative approach is to pretrain the base network via self-supervised learning (which does not require labels). The basic results of my work are: i) in the low labelled data regime, self-supervised pretraining is superior to supervised pretraining; ii) self-supervised pretraining a *second* time on a large labelled dataset from the target dataset (e.g. unlabelled skin lesion images) before fine tuning can yield additional gains. The

thesis work is the basis of the paper: *Self-Supervised Learning for Skin Cancer Diagnosis with Limited Training Data.*

Research interests

AI applications to medicine

Until quite recently I was of the view that the most important area of AI to work on was medical applications. This is still an area of interest, and I hope to contribute in the future, particularly to AI applications to human longevity (shockingly still an underfunded area compared to other areas of medicine).[1] My general view on this now though, is that the biggest breakthroughs (from the machine learning point of view) will come from training large foundation models, e.g. AlphaFold. There are also a large number of people working on AI applications to medicine.

Human level AI may arise in the near term. This may pose an existential risk, according to AI researchers

We can define human level AI as an AI system that can reproduce all human cognitive ability, including abstract reasoning, detection of emotional valence of faces, etc. Once a system is human level, it will likely be beyond human level in a short period of time, by simple scaling arguments (and scaling has more than shown its power of late c.f. GPT-x). There is substantial disagreement among AI researchers around both the timeframes within which human level AI may arise; and whether or not this poses an existential risk. However, by some accounts the median estimate is 36 years to human level AI.[2] A smaller but non trivial number believe this would pose a 5-10% risk of human extinction. If we take these probabilities as the ground truth, i.e. there really is a 5% risk of human extinction, this is extremely concerning. Furthermore, the true probability of extinction may be substantially higher. For one, it is unclear how familiar the mean AI researcher is with the AI safety literature. Moreover, experts are not always sanguine in timeline estimates. Consider for example the median 'expert' estimate of timeframe to develop airplanes in early 1903. Similar statements can be made about expert opinions on the viability of splitting the atom or building nuclear weapons.

The Turing award is considered the Nobel prize for computer science. In 2018 three individuals were awarded this prize for work on deep learning, and it is not unreasonable to view these people as 3 of the most expert AI researchers in the world. Of these three, two believe AI may pose an existential risk (Bengio and Hinton) with the third (LeCun), disbelieving this claim. Hinton has been coy in giving exact probabilities, but his interviews make clear he believes the risk is not low and may be in the realm of 50% or higher. Naively extrapolating a 50% extinction risk view from Hinton and Bengio, 0% for LeCun, and assigning a $\frac{1}{3}$ weight to each expert opinion gives a $\frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$ extinction risk. The point of this is not to claim we have positive knowledge of the probabilities - Hinton himself does not claim so - but that the probability of extinction is *plausibly not low.*

IGNORE

Different terminology we can use (i.e. buzzwords…)
- Reward misspecification
- Reward hacking
- Specification gaming

Could also give robot grasping example!

END IGNORE

How might AI pose an existential risk?

There are various ways beyond human level AI may pose an existential risk, with the most obvious being a mis-alignment between the AIs values and humans values, broadly construed. This can occur for at least two reasons: i) Through failure to specify the right goal, the paradigm example being a paperclip maximiser.[3] ii) Through failure during training to *learn* the intended

goal, and instead to learn a different objective not intended by the developers. Such a system is called a mesa-optimiser.[4] A simple example is: a system is trained to reach the door in a maze. On the training distribution, all doors are red. On the test distribution, the doors are all blue, and there are other (non-door objects) that are red. Although the performance at training time is ostensibly aligned, the alignment is spurious and at test time the system goes to red objects instead of doors. In other words, the AI learned the wrong objective during training. Specifying the correct goal as in i) Is called outer alignment. Ensuring the AI learns the intended goal during training as in ii) is called inner alignment. Both are unsolved problems at present. A potential solution to ii) is to ensure an AI is always truthful, although this also has subtle failure modes associated. For example an AI may be truthful at training but not test time. The inner/ outer distinction is due to Hubinger et al. (2019).

Few people are working on this problem

We can surmise that: *if an AI is sufficiently capable, and has misaligned objectives, it will pose an existential risk*. This is an area that: a) few people are working on, compared to e.g. capability research or narrow AI applications; b) it appears possible for individuals or small groups (with limited compute budgets) to make outsized progress through both conceptual clarifications and empirical work. Some examples include: Hubinger et al. (2019) conceptual work on clarifying inner vs outer alignment (discussed above); work by Olah et al. on mechanistic interpretability of transformers;[5] work by Burns et al. (2022) on unsupervised truth extraction from language models.[6]

Tentative research agenda

I hope to contribute on two fronts: i) Conceptual clarification of failure modes for AI alignment. e.g. inner vs outer alignment failure. ii) Empirical work aligning current AI systems, which will hopefully generalise to future systems e.g. extracting truth from large language models. As a first step I am presently studying the AI alignment literature.

References

[1] https://www.lifeextension.com/magazine/2011/9/why-arent-more-wealthy-people-funding-aging-research

[2] Survey conducted by MIRI in 2022 https://wiki.aiimpacts.org/doku.php?id=ai_timelines:predictions_of_human-level_ai_timelines:ai_timeline_surveys:2022_expert_survey_on_progress_in_ai

[3]*See e.g. Bostrom, N. (2014). Superintelligence.*

[4] Hubinger et al. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems

[5] See work on transformer circuits e.g. https://distill.pub/2020/circuits/ and https://transformer-circuits.pub/

[6] Burns et al. (2022) Discovering Latent Knowledge in Language Models Without Supervision