Research Statement

The intersection of cognition, computation, and neuroscience has always fascinated me. I've made empirical and theoretical contributions to many topics, including analogy, heuristic use, exploratory behaviour, neuroeconomics, and reinforcement learning, and have found support for novel forms of learning in humans and machines. Many of these contributions were ahead of their time. For example, my first publication from my undergraduate honours project was PageRank for the mind three years before Google's PageRank (see PageRank wiki). My long-standing research focus is how people learn categories from examples.

More recently, my contributions have shifted toward machine learning and artificial intelligence. I conduct large-scale analyses of human behaviour in naturalistic settings, derive embedding from human judgments that are orders of magnitude larger than existing approaches, extend deep learning approaches to be more human-aligned in terms of behaviour and accounting for brain activity, and develop tools to accelerate scientific discovery, such as in the BrainGPT.org project. Below I outline my journey from psychology to model-based neuroscience to my current efforts.

I end by discussing BrainGPT, which speaks to my vision for Generative AI. I believe scientific discovery will be increasingly automated and human-machine teams will be the norm. The plan is to first apply these methods to neuroscience, followed by other knowledge-intensive fields and translational impact (e.g., drug discovery).

Past work in Category Learning

Concept learning draws on several interesting component processes associated with attention, object recognition, memory, goals, and generalisation. I found this topic a useful entry point to addressing fundamental questions, such as how people behave flexibly and how the brain realises the mind.

When I entered the field, dominant category learning theories were rigid and insensitive to a learner's goals. For example, prototype models always represent all category experiences with a single node in memory, whereas exemplar models always store each experience as a separate node. With colleagues, I developed a learning model that moves between these extremes depending on the learning problem and learner's goals. The model clusters together related experiences in memory until a goal failure triggers storage of the surprising experience, which itself can later be abstracted by subsequent similar experiences. The model is very successful at capturing a range of learning and memory phenomena and is still the best model of human learning more than two decades later.

While simple and understandable, the model's components can be related to processes involved in top-down attention, knowledge representation, recognition, familiarity, and error correction. Quantifying these cognitive processes provides a means to understand their brain basis. Early on, I grounded the model mechanisms in the Medial Temporal Lobes (MTL) and medial prefrontal cortex (mPFC), focusing on the hippocampus, which was strangely neglected in the cognitive neuroscience of category learning literature prior to my work. When the timing was right, I tested these predictions in model-based fMRI, which successfully characterised the function of the MTL during category learning. In our first paper, across learning trials, we found the MTL's activation profile closely matched the model's familiarity/recognition strength signal, whereas the error-correction signal matched MTL activity at corrective feedback. This was the initial step down a path that led to understanding how the hippocampus and mPFC interact to support goal-direct concept learning. While studying these systems, we pioneered and systematised a number of approaches in model-based fMRI, such as extending brain decoding approaches for stimuli (e.g., is the participant viewing a house or face) to model states (e.g., does brain activity better predict the internal state of model A or model B). In doing so, we highlighted the importance of evaluating competing neurocomputational accounts rather than assuming one's preferred approach is correct, which is important because a model-based analysis is only as good as the model used.

More recently, we established how the hippocampus and medial prefrontal cortex (mPFC) interact to support goal-directed learning. We used pattern-similarity analysis to uncover non-spatial

multidimensional representations in the anterior hippocampus that are modulated by task goals through interactions with mPFC. Task goals were reflected in attention weights learned by a cognitive model fit to behaviour, which we later found reflected individual differences in both behaviour and BOLD response. These attention weights also indexed the intrinsic dimensionality of visual representations in the ventral stream, which we established by applying a dimensionality measurement technique we developed. We discovered that the mPFC learns a goal-relevant compression code that accentuates relevant information, which helped explain activity in both the anterior hippocampus and the ventral stream. We followed up with a theory and simulations in which this same non-spatial learning mechanism captured place and grid cell activity, offering a new view on grid response. Our most recent deep learning work in computational neuroscience goes further to question the need for intuitive cell types and built-in attentional mechanisms to account for brain function.

Foundational Contributions

I haven't shied away from addressing foundational issues in cognitive science, neuroscience, and machine learning. For example, we questioned the restrictive approach taken by rational Bayesian approaches that were dominant in the 2000s. As we noted, these (then ascendant) approaches relegated all of neuroscience and development to tertiary roles in understanding cognition. I believe our efforts played a role in reshaping the research agenda and furthering the integration of cognition, computation, and neuroscience. Here are three example papers that follow our own recommendations on Bayesian modelling. Related, I edited a special issue on model-based cognitive neuroscience that broadened participation in this arena. In pursuing integrative theories of cognition, brain, and computation, I have tried to help clarify thinking on basic issues such as levels of analysis, reduction, emergence, and incoherent notions of biological plausibility, suggesting that the field should instead clearly state claims that can be evaluated by model comparison. In this spirit, we have considered which approaches, ranging from simple models to complex deep learning models, are most consistent with the neural code given the successes of fMRI in uncovering representational spaces despite its limited temporal and spatial resolution. In machine learning, we have considered basic issues, such as what constitutes a short-cut in the covariate shift problem and how to ameliorate it through a "too-good-to-be-true prior."

My lab's policy is to make both our data and code available without restrictions. When resources permit, we make code available in multiple languages with unit testing, proper documentation, etc. to promote uptake, as we did in the dimensionality estimation project using GitHub and Travis CI. We make regular use of open datasets and support open standards, such as Brain Imaging Data Structure (BIDS). Most of our research is supported by open-source tools, and we seek to minimise reliance on proprietary solutions. We are proponents of preprints to speed the dissemination of scientific findings.

Looking Forward

My experiences have positioned me to take advantage of recent advances in machine learning, computing methods, and hardware, as well as the increased availability of large, open datasets. I see opportunities in the Neuro-AI space for encompassing models of brain function that scale. It is now possible to address the neural bases of complex behaviours involving naturalistic stimuli (e.g., photographs, movies, etc.). Rather than disconnected models of the hippocampus, prefrontal cortex, the ventral stream, and multimodal association areas, it will soon be possible to have integrative models that address all these regions simultaneously. My vision is that these encompassing approaches can be expressed in general terms (i.e., coarse grained) to offer general accounts of brain function, or, alternatively when desirable, further decomposed down to the level of a neuron, all the while accounting for behaviour.

The ultimate payoff is more robust and encompassing models that can illuminate the function of the healthy brain, as well as the compromised brain, whether the insult is from disease, stroke, or trauma. The methods we are developing for evaluating models are directly applicable to developing brain machine interfaces (BMI) that can address higher-level thought far away from the sensory-motor periphery. Data science approaches will be key to transform large real-world datasets of human

behaviour into models of memory, learning, knowledge representation, and decision making to advance neuroscience through rich model-based analyses.

As discussed, cognitive models can be very useful in simultaneously addressing brain and behaviour through model-based analyses. However, cognitive models suffer from some key limitations. While cognitive models are useful in identifying neural correlates suggestive of function, there remains a gap between the higher-level constructs in these models and neurons. This gap limits the applicability of models and leaves key questions, such as how does the brain make a symbol, unanswered.

For example, while Rob Mok and I successfully related representational units in our cognitive model with place and concept cells in the hippocampus, an explanatory gap remained. Whereas our cognitive model had a small number of representational units (e.g., clusters that capture related experiences), the hippocampus has many cells. In current work, we decompose these higher-level cognitive constructs into neuron-like computing elements while still capturing the input-output characteristics of the abstract model. We introduce the notion of "neural flocking", akin to how birds or artificial life agents flock according to local rules absent a central controller. We hypothesise this flocking behaviour is implemented by recurrence in the hippocampus, whether it be internal or so called big-loop recurrence. The upshot is that the aggregate (i.e., the flocks) of this lower-level model matches the higher-level model (i.e., the clusters), which provides a multi-level explanation of processing that spans behaviour to neurons. I will pursue this "levels of mechanism" approach in other endeavours in which higher-level models, which account for behaviour, can have a component decomposed into its own mechanism when it is desirable to make closer contact with fine-grain measures and phenomena.

Making deep learning more human aligned

One limitation of cognitive models is that they are restricted to processing stimuli handcrafted by the experimenter and cannot be directly applied to naturalistic stimuli. Modern deep learning models address this issue. In addition to being able to process naturalistic stimuli such as videos, deep learning models offer additional advantages such as multiple processing stages or layers that can be put in correspondence with multiple brain regions to offer more encompassing theories of brain function. One future research direction is extending these approaches to be more brain-like. Early efforts along these lines have incorporated goal-directed, top-down, attention mechanisms and evaluated their efficacy on large, naturalistic datasets. We have also incorporated generative replay into deep learning models to better understand its role in knowledge consolidation and address limitations in existing machine learning approaches.

One overarching aim is integrate deep learning approaches with cognitive models, such as our own formal accounts of how mPFC and the hippocampus support category learning. The payoff of this approach is that we can offer accounts of how multiple brain regions interact to support rapid learning and generalisation based on only a few training examples. Rather than offer a model of a single narrow task or a specific brain region, this comprehensive approach addresses interesting human behaviours at scale. We are revisiting our work on attention in deep learning models following this approach. One idea we are evaluating is that the information needs of the hippocampus and mPFC direct encoding along the ventral visual stream. We will test a general theory of coordination across brain regions, which we cheekily call "the costly energy principle" in which there is a controller (e.g., mPFC) and peripherals (e.g., ventral stream regions) that aim to reduce their activity while not disrupting processing in the controller. The overarching principle is that the brain prefers to preserve computing resources, when possible, such that cells not relevant to higher-level computations can disengage. We will evaluate this controller-peripheral approach as a general account of brain function and quasi-hierarchical control. Our approach will lead to machine learning models that use minimal resources and are capable of continual learning.

Properly evaluating deep learning models is a challenge for neuroscience. Currently, the field relies on establishing correspondences between model layers and brain regions based on shared variance. While these methods superficially differ (e.g., encoding approaches, RSA, CCA, CKA, etc.), they all share the implicit assumption that correlation implies correspondence. However, not all variance in brain measures is of interest or even task related. We propose a stronger test of correspondence based on substitution: If a model layer corresponds to a brain region, then replacing that layer with brain activity

should drive the model's activity toward an appropriate output. We applied this approach to object recognition using a simple linear mapping or translation from brain to model space and found, contrary to the zeitgeist of the field, that all regions along the ventral visual stream best corresponded to late model layers after 60ms. This result likely arises from long-range recurrent connections from “late” visual regions, such as inferotemporal cortex, to “earlier” regions. Our approach to model evaluation has the pleasant side effect of being applicable to brain machine interfaces – we found that only 10ms of multi-unit recording data from monkeys is needed to drive the deep learning model. With human fMRI, zero-shot decoding was successful for object categories held out from training, indicating the generality of brain-model mapping.

BrainGPT: A tool to accelerate neuroscience research (and other knowledge intensive fields)

I was attracted to computational modelling in part because it can effectively compress or summarise a vast empirical literature. A good model can explain hundreds of empirical studies consistent with it. However, given the exponential growth of the scientific literature, even with models, the literature is becoming fragmented, and it is impossible for scientists to keep up. Potentially disruptive findings are overlooked due to the rapid expansion of the scientific literature. The challenge of integrating findings may exceed human abilities. Already, specialist solutions have been developed to address important scientific questions in protein folding, drug discovery, and materials science. Rather than replace humans, I foresee a human-machine teaming solution using large language models (LLMs). While my initial efforts will focus on neuroscience, I intend for the approach to apply broadly and will promote its adoption across knowledge-intensive endeavours.

One drawback of LLMs is that they can confabulate (i.e., “hallucinate”), which makes them problematic for information retrieval. These models cannot be trusted to correctly summarize the scientific literature (cf. Meta's Galactica). These failures are really features of the generative model – LLMs aren't knowledge graphs (e.g., Wikidata) that store facts. Instead, LLMs perform a synthesis by integrating large quantities of noisy and imperfect information. Likewise, the scientific literature is noisy, sometimes contradictory, and doesn't always replicate. In the BrainGPT project, we cater to LLMs' strengths and avoid their weaknesses. Instead of being used for information retrieval, we draw novel inferences from a large pool of data where “mixing” of facts is desirable because the goal is to generalize and predict unknown outcomes, such as predicting how different neural measures relate and the statistical power of candidate study designs.

We have extended LLMs by providing additional training in the neuroscience domain using LoRA. BrainGPT (an open weight LLM +LoRA) will serve as a generative model of the scientific literature, allowing researchers to propose study designs as prompts for which BrainGPT would generate likely data patterns reflecting its current synthesis of the scientific literature. Modelers will use BrainGPT to assess their models against the field's general understanding of a domain (e.g., instant meta-analysis). BrainGPT could help identify anomalous findings, whether because they point to a breakthrough or contain an error. BrainGPT will also illuminate the basic structure of fields by tying changes in benchmark performance to the training curriculum. For example, in what situations do behavioural studies help constrain predictions about brain activity? Using BrainGPT, we can also better quantify citation biases, evaluate which aspects of the literature are most reliable and replicable.

To evaluate BrainGPT and other LLMs, we have developed a novel forward-looking benchmark with the help of 75+ neuroscientists from the BrainGPT.org community (2440+ neuroscientists). The benchmark, BrainBench, assess the ability to predict neuroscience results from methods. LLMs predict neuroscience results from methods better (85% vs. 63%) than neuroscience experts. Accuracy improves further when the judgments and confidences of human experts and LLMs are integrated in a Bayesian fashion. Our findings should change the course of the field, ushering in a new era of human-machine teaming in the pursuit of scientific discovery. With recent support from Microsoft Research, we have trained LLMs from scratch rather than relying on existing solutions. This approach gives us full control of the model and allows us to better understand the bases for its performance. All materials and models are open source. One future objective is to develop tools that use BrainGPT to determine which experiment should be conducted next. Further into the future, we will train models on raw data in addition to text and figures from articles.