Motivation for the Research

Paula Parpart

1 Learning and Decision-Making in Humans and Artificial Agents

My current research looks at the computational mechanisms underpinning learning and decision-making in humans and artificial agents. In particular, I am interested in the role that flexible generalization with simple cognitive models plays in both human and machine cognition.

1.1 Generalization in Humans and Machines

In 2017, Lake, Ullman, Tenenbaum and Gershman (2017) argued that despite the major advances in artificial intelligence and deep neural networks reaching human-level performance on many tasks, these systems still differ from human intelligence in significant ways as they are missing key ingredients from human cognition such as 1) building causal models of the world that help understanding and explanation, 2) intuitive theories of physics and psychology to embed the knowledge learned 3) compositional learning and 4) learning-to-learn to generalize knowledge to new tasks and environments. This has led some researchers to argue that reverse-engineering human solutions to difficult computational problems is the way forward. Progress was made on building more human-like agents that are able to address the missing cognitive ingredients. For instance, Hamrick et al. (2018) and Battaglia et al. (2018) demonstrated that implementing a relational inductive bias within deep learning architectures can facilitate learning about entities and relations in a simple physics task, and crucially, support combinatorial generalization.

Generalizing beyond one's experience remains one of the key hallmarks of human intelligence that is still out of reach for current deep learning approaches (Battaglia et al., 2018). Even with the impressive increase in large language models' (LLMs) capability to generate text from billions of data points in recent years, they still struggle to handle completely new information, especially in one-shot learning scenarios. This is because large language models (e.g., GPT-4) tend to exhibit strong in-distribution generalization but often struggle with out-of-distribution generalization. The artificial intelligence research community is treating topics such as one-shot learning, transfer learning, and compositional generalization as a priority in order to achieve more human-level inference capabilities or inch closer to artificial general intelligence (AGI) (Wang, Jindong, et al., 2022).

In particular, I am interested in the role of heuristics (i.e., regularizers) for generalization in humans and machines. I propose that a key aspect that sets apart human generalization from artificial agents is the extraction of simple heuristics (strong inductive biases) in different domains. From previous work outlined above (Parpart, Jones & Love, 2018), we know that heuristics correspond to strong inductive biases that can be encoded as the prior of a Bayesian model. Several decades of research into human decision-making and learning also suggest that heuristics are an indispensable part of human cognition being able to flexibly generalize to new environments representing more biased solutions that overfit less (Bramley, Dayan, Griffiths & Lagnado, 2017; Czerlinski, Gigerenzer & Goldstein, 1999; Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier, 2011; Pitt & Myung, 2002). Therefore, the kinds of questions that I ask in my research are **How do we choose the best decision-making strategy to generalize to novel situations? Can we build an artificial agent capable of learning simple heuristics (regularizers) from data?**

One way to address these research questions is via a learning-to-learn approach. *Meta-learning* has established itself as one of the most successful approaches to improve one-shot generalization in recent years (Ortega et al., 2019; X Wang et al., 2016): the aim is to *learn the learning* procedure from the data itself, and produce flexible, data-efficient learning systems through the acquisition of inductive biases from data (Bengio, Bengio & Cloutier, 1990; Ortega et al., 2019). Rather than having biases built into the system, in meta-learning, these biases are learned by training on a distribution over tasks (e.g., learning not only to find rewards in a single maze, but training on a broad range of mazes which ends up learning a more general-purpose strategy for exploring new mazes). When agents are trained in this way, they are able to extract task structure in the task distribution that allows them to generalize to new tasks. This approach has been successfully used to exhibit causal reasoning in artificial agents. Dasgupta et al. (2019) trained a recurrent neural network with model-free reinforcement learning on a range of problems that each contained causal structure. Importantly, the agent was able to extract causal knowledge from the distribution of tasks, perform informative interventions and make counterfactual predictions without being explicitly given any formal principles of causal inference. In my work

I ask whether, in the same way as the learning-to-learn approach was used for causal inference, can we build a meta-learning agent capable of discovering heuristics and that uses heuristics for generalization?

In this project with Chris Summerfield at the University of Oxford, we worked on precisely these questions, and I will outline what we found below. We built a meta-learning neural network architecture with the goal of learning the optimal heuristic (or other decision-making strategy) from scratch, simply by showing data of a particular statistical structure to the artificial neural network. Importantly, the optimal strategy discovered is not any optimal heuristic - it is the heuristic that best generalizes to new, unseen data. We set up the neural net architecture to inherently include cross-validation as part of its architecture to measure generalization performance. This means our meta-learning neural net learns the best possible regulariser for each statistical environment. Subsequently, we generated tons of synthetic data with a structured generative process. This generative process allows experimental manipulation of the statistics of the environment, (e.g., number of features, feature covariance, weight distribution of features, and crucially how noisy the data is). In this way, the input space becomes the different types of statistical environments, while the output space are the different generalization strategies. In our framework, heuristics correspond to a transformation of the observable weights (i.e., ordinary least squares) in a dataset. More precisely, it is a sigmoidal transformation of the weights resulting in a compressive nonlinearity (i.e., this leads to binarization of the weights in the case of tallying heuristic, which uses only unit weights, and putting various weights to zero in the case of the Take-The-Best heuristic, which relies only on the most predictive feature). Importantly, the functional forms of the heuristics can be learned with stochastic gradient descent. A key question that our simulations with the meta-learner were able to ask is: Are there situations for which these compressive nonlinearities (heuristic models) have better generalisation performance than the ground truth generative weights?

We find that 1) In various datasets for which heuristics have been shown to perform well, these can be learned from scratch with our meta-learning agent. For instance, the meta-learning agent discovers the Take-The-Best heuristic as optimal when statistical environments had more uneven weight distributions or higher feature covariance, while Tallying tends to be discovered when environments had more evenly weighted features, in line with previous literature. 2) The meta-learning agent discovers heuristic models as optimal solutions when there is 'early' noise present (i.e., noise on the input features themselves), representing a noisy encoding of the decision stimuli in human learners.

This finding is interesting as it directly links to a series of existing studies showing that finite computational precision can result in compressive nonlinearities representing robust strategies in the face of early decision noise, ranging from domains such as economic choice to decisions about percepts (Juechems, Balaguer, Spitzer & Summerfield, 2021; Summerfield & Parpart, 2022; Summerfield & Tsetsos, 2015). Hence, a question that our work poses is that, assuming these heuristics represent the true decision-making processes, do humans act as if their internal representation of the predictive value of the inputs is distorted away from their true value (e.g., compressions)? Why would decision-making have evolved in this way? One possible explanation is that in the face of selective pressure for generalization-maximising strategies, and assuming finite computational precision, i.e., noise at the decision stage, the policy (strategy) that maximises generalization performance on novel data is indeed distorted away from the ground truth weights, making it the most robust strategies. This would also indicate that previous arguments for why people might use compressive models based on the bias-variance tradeoff (simpler models being more biased) may not be needed, if indeed heuristics result from decision-making with finite computational precision alone.

In our *Annual Review of Psychology* article (Summerfield & Parpart, 2022), we go into more detail on this argument, bringing together literatures from both perception, decision-making, and economic choice that have previously been distinct. We explain the meaning of 'early' and 'late' noise, and show that the same common principles, i.e., normative principles from efficient coding and Bayesian inference, can help explain why it is adaptive for people to rely on compressive discretization heuristics, but also why sensory illusions, choice history biases, decision biases, framing effects, nonlinear utility functions etc. appear. We show how these seemingly irrational behaviours can be seen as the result of the same distorted encoding functions at the neuron population level, which are adaptive for agents with finite computational resources (i.e., efficient coding).

Further impact of this research is that our meta-learner acts as a tool for discovering regularisation algorithms for any type of dataset in machine learning. That is, it also becomes possible to discover solutions that move beyond the already extant regularisers such as ridge regression or lasso regression. In our work, we identify

several statistical environments where a solution is discovered that can be formalised as a novel heuristic (or regularizer) and outperforms existing ones. To highlight the impact for machine learning, we have also started to used our framework to model canonical AI datasets, such as the image classification MNIST dataset, and find that it performs well in such classification tasks. I would like to extend this research by applying it to other canonical AI datasets to see if our meta-leaner can discover optimal strategies in various contexts that go beyond existing approaches.

In my future research, I plan to take the research on *Generalization in Humans and Machines* and develop it into a bigger research agenda. I plan to ask research questions such as **How can simple heuristics be used by advanced AI systems for decision-making that is more flexible and human-like? Could we use the AI to discover robust heuristics as decision-making strategies in domains such as healthcare?** For instance, in the domain of healthcare there is already a body of work showing that heuristics are widely used, particularly in clinical decision-making by doctors and medical staff, but also in areas like research or diagnosis. They help clinicians make rapid judgments and decisions, especially in time-constrained situations such as the ER. Our approach would be able to test the accuracy and generalization ability of various existing medical heuristics, as well as discover new ones.