10

Conclusion

In 1955, Herb Simon put forth the challenge facing more realistic theories of human intelligence:"Broadly stated, the task is to replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist." This thesis hopes to do exactly that. By taking into account the circumstances under which intelligent behavior manifests – both the limitations on resources, structure in the environment, and how these two interact – I provide new computational models of human probabilistic inference, that are psychologically plausible. Without plausible algorithmic solutions to rational or normative inference in structured Bayesian models, they remain unsatisfying as models of human cognition. We also cannot leverage their many desirable properties in building intelligent machines. The ideas furthered in this thesis, of leveraging environmental structure via flexible re-use of previous computations to simplify inference, bring Bayesian models of intelligent behavior back into business.

Further, these models parsimoniously explain a wide range of empirical findings about non-normative inference, and how humans can sometimes be so close to optimal, and at other times (with the same cognitive resources), so biased – and biased in so many different context-sensitive ways. These insights also lead to entirely new ways to understand and engineer artificial systems, via manipulation of the environments in which they learn and function. This confluence suggests links between the analysis of ecological rationality in humans and in machines, leading to new lines of research into understanding both.

Open Questions and Future Work

Model acquisition An important question not addressed directly in this thesis is of how structured probabilistic models are acquired in the first place. In our studies of human cognition, we have distinguished between 'learning about the world' or 'potential knowledge', and 'learning to think' or 'realized knowledge'. Most of the work on human cognition presented here operates solely on the second, i.e. in the realm of internal processes within the mind, *after* all external knowledge has already been gained and represented as a probabilistic model. In some of these examples, we verbally provide the data generating process, i.e. the underlying structured probabilistic model (for example the urn experiments in Chapter 7), but in several others, we assume this is known from pre-experimental experience (for example in the scene statistics domain used in Chapters 5 - 7). How might these structured Bayesian models be acquired via interaction with the environment?

One way to look at model acquisition is as a higher level probabilistic inference. That is, the representation we acquire of how a domain works is by searching over some space of possible models built from structured primitives, assigning probabilities for how well they explain the observed data (the likelihood of that model), and then choosing a model such that it has high posterior probability. This is of course, also a very challenging inference problem [402,46]. Further, it risks passing the buck further down: how do we know what the primitives to building a good hypothesis space of models is? Several findings show that many primitives of structure might be innate and available at birth before any interaction with the environment [431,60], but that too leaves open the question of how (potentially via evolution) such primitives came to be innately encoded. The challenges of how this search over models might occur, therefore,

forms another key criticism of structured Bayesian models for intelligent behavior. These criticisms have been developed further largely by the 'connectionist' approach to cognition[377,299] that posits instead that structure emerges from interactions with the environment, via low-level learning mechanisms, rather than via a discrete search over the space of possible structured models. Most modern approaches to artificial intelligence, i.e. deep neural networks, follow in this tradition. Recent work has also suggested that more structured forms of intelligent behavior can in fact emerge from such 'low-level' learning[75,472,42].

Explicitly structured probabilistic models however have several desirable features – like efficient learning[238], greater generalizability[261], and an accurate representation of uncertainty[190]. An important direction of future research therefore is to find ways to harness these advantages while avoiding the prohibitive costs and implausibility of learning these models via search. One possibility is that structure in the environment (and amortization procedures that reflect this structure) can alleviate these costs, in the same way this thesis suggests that it could alleviate the intractability of inference within a learned model. Models for amortized inference, including the inference network studied in Chapter 7, have striking similarities to connectionist models. This suggests future directions of research that explore this connection, to build new hybrid models that combine the complementary advantages of these approaches.

Two kinds of learning Another interesting direction is to consider the interaction between learning about the world, and learning to make inferences in it. Although we have so far treated these entirely separately (in this thesis as well as in the previous section), in most real world domains, these are not separate tasks. In fact, we almost never learn models directly, we learn them as an intermediary towards performing some task that requires an inference within that model.

As an illustrative example, we consider a classic example from reinforcement learning, of latent learning in Tolman's rat mazes[454]. Here, rats learned to navigate mazes of very specific shapes, to get to a reward. Simply memorizing the actions required to get to the reward in these mazes would have been sufficient to always receive the reward. Tolman[454] found however that rats developed a more abstract model, or 'cognitive map' of the spatial position of the reward with respect to their starting point. This was evidenced by the fact that the rats find the reward by navigating directly to it, in close to a straight line, when the walls

of the maze are removed*. In this case, learning the model is like learning the spatial position of the reward with respect to you. This captures something about the underlying structure of the environment (spatial in this case), and this knowledge can generalize to give reasonable performance in different situations – like starting from a different initial points, differences in the structure of the maze, or obstacles in the way. Inference in this model corresponds to planning ones actions (within constraints like walls of mazes, and wanting to minimize energy spent) in order to get to the reward. The task that the rats are trained on only really requires the ability to make some very specific inferences in an otherwise much more general / complex model, with no explicit requirement to represent an intermediate structured model. But we see that they acquire such a representation nonetheless. In other words, they could learn a purely *discriminative* model (for the purposes of the task they are trained on), but instead learn an at least partially *generative* model. This is characteristic of several domains – even when abstract models are useful, they are rarely explicitly taught or tested. Rather, they are acquired as an implicit intermediary to a task that additionally also requires inference in such a model. In this work, we assume that the model is learned, and the only remaining challenge is in the inference. Future work should consider the problem of jointly learning a structured model for the environment, and learning to perform efficient inferences in this model.

Shaping our environments So far, we have assumed that the interaction between the environment and the intelligent systems that live and learn in it is one-directional – we have only considered the impact of the environment on the procedures learned by agents that interact with it. However, intelligent agents frequently influence and even entirely form their own environments. This two-way interaction is especially pertinent in domains like language where the production mechanisms themselves are shaped and limited by human cognitive abilities. Our ability to learn and understand compositionally structured languages is learned from data produced by other humans' ability to produce these compositionally structured languages. The role of shaping one's environment is also relevant in other domains that are not as directly produced by humans. As far back 1956 in the study of category and concept learning, Bruner[50] presented a distinction between learning through passive reception of observations and through active se-

*The actual experiment did not remove the walls of the maze but replaced it with a maze that contained several radial arms, and found that rats take close to the shortest path to the reward by choosing the right arm.

lection of observations in support of hypothesis testing. Much subsequent work has expanded upon the significant impact that active information seeking behaviors have on learning[322,295,189], suggesting that even very young children can and do engage in behaviors that shape their own learning environments[381,170,316].

A 'rational analysis' approach to active learning posits that humans maximize information gain, subject to the costs of information gathering. However, similar to our quandaries about exact probabilistic inference in humans, exactly computing information gain is nearly intractable. Further, several studies often find biases in people's information seeking tendencies[227,21]. Future work should consider how processes like caching, re-use and amortization that can ease the computational burden of normative information seeking, and potentially lead to new rational process models of active learning.

Grounding the theory A key aspect of this thesis is to more explicitly consider the role of memory in human inference. Even within this framework of using memory as a computational resource, it is yet to be understood what the contributions of different memory mechanisms (episodic, semantic, procedural, etc.) might be. This thesis has been largely agnostic to the specific kinds of re-use and how they might be realized in human memory systems. Future work can more explicitly investigate these different kinds of re-use. In the same vein, ecological rationality and biased judgments have been studied extensively, and several models for these have been proposed. We have discussed some of these alternatives in this thesis, as well as how many of them fit into the broader framework of amortized inference. Future work can work towards better understanding how these many models and different memory mechanisms fit together and inform each other.

Another consideration is how such theories might be implemented in the brain. Future work should look for signatures of amortization in the brain, and better understand which parts of the brain are involved in the different kinds of learning discussed here. We have briefly discussed the neural plausibility of approximation algorithms like Markov chain Monte Carlo, and variational inference in Chapter 3. A main proposal of this thesis is a hybrid model that incorporates aspects of both algorithms. Future work should consider how such hybrids might be realized in networks of neurons.

Closing thoughts Amortization as an approach to ecological rationality also has much broader, and further-reaching implications for cognitive science, beyond the topics studied in this thesis. A better understanding of the underlying computational principles of ecological rationality can shed light on one of the longest-standing debates on the basis of human cognition: the conflict between compositional structure and simple statistics, in models of cognition.

Systematicity and compositionality[120] in structured representations permit the kinds of flexible generalization, far beyond direct experiences, that humans commonly exhibit.[181,408,484,401] This flexibility however comes at a cost. While systematic and compositional representations allow recombination of its components in many different ways to provide solutions to new problems, inferring the solution to a particular given problem – by inferring the right combination of components in this large space – is very expensive. In other words, these models are 'generative', having an explicit representation of the underlying generative process that produced the observed data, from which responses to specific queries are yet to be computed (at possibly very high computational cost).

Statistical approaches on the other hand, do not invoke an intermediate generative model. They instead directly learn to provide responses to queries. Without access to an explicit generative model, we lose the potential to generalize flexibly beyond direct experience. However, 'making an inference' is no longer a challenge, since these approaches directly provide this response. In other words, statistical models are usually 'discriminative'. They do not separately represent the underlying data generating process, and instead directly model the mapping between observations and response.

While structured generative models give very good generalization, inference in them is often intractable. Statistical discriminative models on the other hand are poor at generalization but can make fast, often heuristic, inferences. Each of these therefore have been evoked to model different aspects of human cognition across several fields including word semantics[439,377,162], probabilistic judgment[331,458], concept learning and categorization[50,302,408], and reinforcement learning[145,76,251].

A crucial observation however is that these two possibilities simply populate the far ends of a spectrum in the trade-off between generalization and tractable inference. While intelligent systems do generalize flexibly, they need not generalize indiscriminately. They should adapt to the environment to choose what

kinds of generalizations are important, and sacrifice other generalizations (in favor of statistical pattern recognition, or memorization) in the interest keeping inference tractable. This allows for intermediate models that lie between the two extremes of entirely compositional representations, and entirely statistical ones. Where on this spectrum is 'optimal' for an environment or domain, will be determined by the ecological distribution of queries we encounter in it. We do not represent the world in its full generality, rather we represent the world conditioned on what we will have to do with that representation, i.e. respond to specific distributions of queries.

This thesis provides a powerful new theory for how such ecological rationality can come about via the amortization of previous computations. This provides a mechanism for learning representations that could trade-off flexible generalization and tractable inference, in a domain-sensitive way. These insights pave the way toward hybrid models that combine the complementary advantages of structured generative models and statistical discriminative models. Not only does this have significant implications for our understanding of human cognition, these insights can also be used to build better, and more human-like, artificial intelligence.