Précis for
Algorithmic approaches to ecological rationality in humans and machines

Ishita Dasgupta
Harvard University

Introduction

In a complex and ever-changing world, how do humans reason as intelligently as they do—especially given limited energy, data, and time? This question is essential to understanding the underlying principles of almost all domains of cognitive science. Theories developed at the computational level of Marr's taxonomy (Marr, 1982) formalize the information processing task required of various cognitive facilities and posit a 'rational solution' to these tasks (Anderson, 1990). Bayesian models provide such a rational solution for how to reason in situations of uncertainty, in particular when information needs to be integrated across different sources. However, computing these responses via exact Bayesian inference is at best expensive, and at worst intractable. Yet empirical findings show that human behavior is often consistent with these rational responses (Griffiths and Tenenbaum, 2006). How might we be computing these difficult inferences, especially within our neural and cognitive limitations? One possibility is that a very efficient and accurate inference engine underlies human cognition. However, in several notable cases, humans display 'cognitive biases'—where their judgments deviate systematically from exact Bayesian inference. These have been the object of extensive study across psychology (Edwards, 1968) and behavioral economics (Tversky and Kahneman, 1974). If, in fact, humans possess an efficient and accurate inference engine, why do they also consistently make these predictable and seemingly irrational errors?

My thesis proposes a unified approach that reconciles these contradictions. The key insight I build upon is that humans are not general purpose computers: we are instead 'ecologically rational', adapting to structure in our environments to make the best use of limited computational resources. I propose models that make explicit claims about how such ecological rationality can be implemented at the level of computational processes, via 'amortization': the adaptive reuse of previous computations. However, amortization can lead to errors when the current query is not representative of past experience. I demonstrate that these errors can explain a vast range of historically observed human cognitive biases, as well as make novel behavioral predictions.

The intractability of exact Bayesian inference also causes it to remain impractical as an

approach to engineering artificial forms of intelligent behavior. Most modern approaches instead utilize more heuristic forms of inference, predominantly neural-network-based function approximation, that are often very different from the provably rational Bayesian approach. This has led to a great deal of consternation regarding the interpretability and predictability of these systems—if they aren't doing the provably rational thing, what exactly are they doing? How can we regulate them? In the second part of my thesis, I examine the role of the environment in these modern machine learning systems. I show how this lens provides new insights into the underlying rules these systems implement, and further, how dependence on the environment can be leveraged to artificially engineer new kinds of intelligent behaviors, such as causal reasoning and compositional language representation.

The central contribution of this thesis is to highlight and formalize the importance of the environment in shaping intelligent behavior. While this concept has been studied (Brunswik, 1943; Simon, 1956; Anderson, 1990; Gigerenzer and Todd, 1999), the primary focus of cognitive science has remained the internal frameworks, mechanisms, and representations within the human mind. In the words of Egon Brunswik, "Psychology has forgotten that it is a science of organism-environment relationships, and has become a science of the organism". In this thesis, I propose models that approximate exact Bayesian inference by adapting to the structure of their environments via amortization. These jointly explain both the remarkable successes of human reasoning (i.e. in making intelligent inferences with limited resources), as well as its seeming failures (i.e. in making predictable judgment errors). I also demonstrate how this principle provides new avenues towards understanding our current artificially intelligent systems, as well as towards building new systems with human-like intelligence.

CHAPTER 2: Approaches to human probability judgment

Chapter 2 provides an overview of previous approaches to probabilistic reasoning in humans. I discuss the 'rational analysis' approach—Bayesian models of cognition—as well as their shortcomings (discussed briefly above). I then discuss bounded rationality—the idea that computing exact normative responses might be outside the scope of the computational resources and psychological mechanisms available at our disposal. I review two main approaches to this: first, rejecting the principle of rational analysis in favor of finding simple but effective heuristics, and second, incorporating constraints into the optimization process. However, computing the boundedly rational response can often be more computationally expensive than the rational solution, raising concerns about their plausibility. I then introduce the framework of 'ecological rationality' and, how its algorithmic realizations (using frameworks developed in chapters 3 & 4) provide a promising way forward.

CHAPTERS 3 – 4: Amortization and approximate inference

These chapters lay out the technical background used throughout this thesis. Chapter 3 discusses the computational challenges underlying exact Bayesian inference, and reviews ways to instead approximate it. Two approaches are discussed—Monte Carlo and variational approximations—along with specific algorithms that implement them. I discuss the challenges of these approaches, their complementary advantages, and ways to combine them. I also discuss the history of their development, their neural plausibility, and previous applications in cognitive science.

Subsequent chapters (5 – 7) present behavioral evidence of both kinds of approximations in human probability judgment and posit that a hybrid model underlies human inference.

Chapter 3 introduces amortization—the adaptive reuse of previous computations—and how this can mitigate computational costs, and give rise to ecologically rationality. Amortization can take various forms within the approximate inference frameworks discussed above. Subsequent chapters (6 & 7) present behavioral evidence of these different forms. I also discuss how modern machine learning (in particular, discriminative methods that encompass the vast majority of modern neural network models) implicitly incorporate amortization, laying the groundwork for later chapters (8 & 9). Finally, I discuss how amortization has historically been implicit in many cognitive theories of probabilistic judgment and reinforcement learning.

CHAPTERS 5 – 7: Ecological rationality in humans

These chapters build and test ecologically rational models of human probabilistic inference, based largely on amortization within approximate inference algorithms. These models parsimoniously replicate a series of historical findings of cognitive biases that often go in opposite directions, and can also explain the context-sensitivity in both the extent and kinds of errors seen, resolving decades of controversy across the behavioral sciences about the rationality of human probability judgment. Further, they make very limited demands on computational resources—the mechanisms implemented are simple and local, and therefore plausible with human cognitive limitations.

CHAPTER 5: Correlated sampling replicates framing effects

The material in this chapter was previously published in Dasgupta et al. (2017a).

Historical findings show that the self-generation of hypotheses in the service of performing probabilistic inference over a large space of such hypotheses produces systematic deviations from rational inference. We consider the example of the 'subadditivity' effect (Fox and Tversky, 1998). If people are told they are in a room with a table in it, they give higher responses to the 'typically unpacked' query: "What is the probability that there is also a chair, a curtain, a computer or any other object starting with the letter C in the room?", than to the 'packed query': "What is the probability that there is also any object starting with the letter C in the room?". More formally, the perceived probability of a hypothesis is higher when the hypothesis is framed as a disjunction of typical component hypotheses. On the other hand, if the disjunction is instead 'atypical' (e.g. "a canoe, a cow, a canon, or any other object starting with the letter C"), people give lower responses, in an effect called 'superadditivity' (Sloman et al., 2004). How can we explain these biases, and their dependence on question framing?

The number of objects that could occur in a room with a table is very large. Exact posterior inference requires enumerating all of these. With limitations on cognitive resources, we assume that people cannot do this exactly, and posit instead that hypotheses are generated stochastically such that the sampled hypotheses form a Markov chain Monte Carlo (MCMC) approximation of the true posterior. The chain is initialized at query-specific information from the framing of the question. Given resource limitations, we assume that people take a small number of samples. While MCMC converges to the true posterior in the limit of infinite samples, in a small sample regime, initialization will strongly influence how many other hypotheses are generated, as well

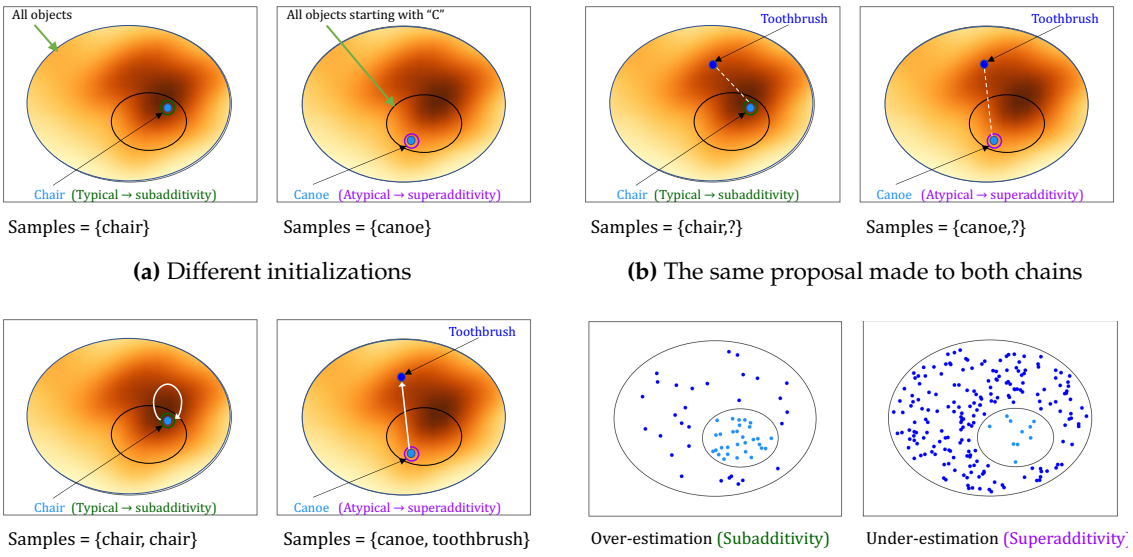


Figure 1: Demonstration of MCMC model. MCMC sampling leads to sub- and super-additivity for different framings of : "In the presence of a table, what is the probability that there is also another object starting with C?". (a) The chain initialized with a typical unpacking starts at 'chair', a high probability hypothesis (darker shading) while the chain initialized with an atypical unpacking starts at 'canoe', a low probability hypothesis (lighter shading). (b) The same (random) proposal of 'toothbrush' is made. (c) Since the probability of 'toothbrush' is higher than 'canoe' the proposal is accepted by the atypically unpacked chain, but since it is lower than 'chair', is rejected by the typically unpacked chain. (d) Typically unpacked chains tarry in the high probability regions of the queried object set, giving subadditivity, whereas the atypically unpacked chain gets derailed into other regions, giving super-additivity.

as which hypotheses are generated. This results in very different sets of samples depending on the question framing, and therefore to different biases in the approximate posterior. An outline of this mechanism is presented in Figure 1. With 7 simulation studies, I show that this model replicates a host of historically observed framing effects, including subadditivity (Fox and Tversky, 1998), superadditivity (Sloman et al., 2004), the weak evidence effect (Fernbach et al., 2011), the dud alternative effect (Windschitl and Chambers, 2004), the self-generation effect (Koriat et al., 1980), the crowd within (Vul and Pashler, 2008) and the anchoring effect (Tversky and Kahneman, 1974; Lieder et al., 2013). The same model also explains why these effects do not manifest in small hypothesis spaces: with the same number of samples in a smaller space, the Markov chain will converge to the true posterior. This chapter also presents 4 new behavioral experiments to confirm the model's prediction that superadditivity and subadditivity can be induced within the same paradigm by manipulating the framing of the query. The model predicts higher biases under cognitive load or time pressure, since these reduce the amount of computation possible, which manifests in our model as a reduced number of samples. These predictions are partially confirmed with novel experiments.

Key features of this model—limited number of samples, and initialization dictated by question framing—can be seen as a rational use of limited resources (Gershman et al., 2015; Lieder and Griffiths, 2019). While taking more samples leads to more accurate probability estimates, the marginal improvement with additional samples reduces with the number already taken. If taking each sample incurs some fixed cost, at some point, a new sample is no longer

worth that cost. Limiting the number of samples taken is therefore computationally rational. Further, the initialization of the model can be seen as ecologically rational. Cues are usually correlated with the relevant hypotheses in the environment (Goldstein and Gigerenzer, 2002). Initializing at these cues (as opposed to randomly) leads on average to faster convergence of the chain, and more accurate probability estimates with the same number of samples—thereby making use of structure in the environment to optimally use limited resources. While this initialization leads to predictable biases in this particular situation, it is likely beneficial on average. An adaptive initialization like this could be amortized, i.e. learned across previous experience. Chapter 7 addresses this possibility more directly.

CHAPTER 6: Humans reuse samples from recent related queries

The material in this chapter was previously published in Dasgupta et al. (2017b) and Dasgupta et al. (2018b).

This chapter explores the implications of amortization in sampling-based inference. For example, the framework in Chapter 5 can answer any of the following questions:

1. What is the probability of a microwave in a room given that I've observed a sink?
2. What is the probability of a toaster given that I've observed a sink and a microwave?
3. What is the probability of a toaster and a microwave given that I've observed a sink?

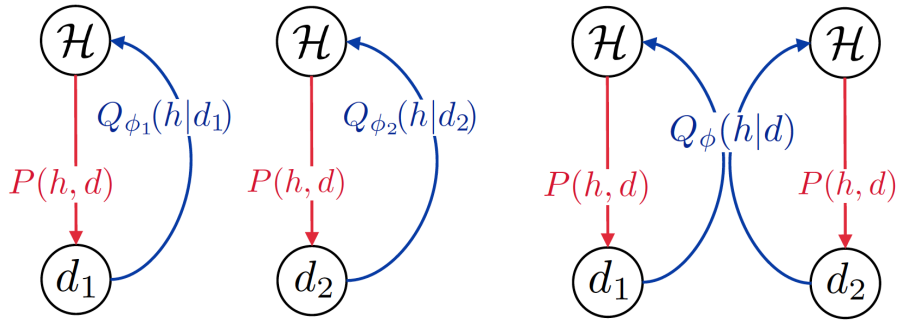
For each of these queries, a new sample-based posterior approximation is formed—the inference engine is memoryless. However, queries are usually not independent. In the example above, the answer to question 3 can be directly obtained from the answers to questions 1 and 2: P (toaster $\cap$ microwave | sink) = P (toaster | sink, microwave) $\times$ P (microwave | sink). Ecological rationality—the utilization of environmental structure to make the best use of limited resources—therefore dictates that we should not neglect all previously computed responses.

This chapter demonstrates (with 3 new experiments) that when sequentially answering two related queries about natural scenes, responses to the second query systematically depend on the structure of the first query, consistent with the adaptive reuse of past computations. New simulations show that different kinds of reuse make divergent behavioral predictions as modulated by a cognitive load manipulation. New experiment test these, showing evidence that people amortize summary statistics of previous inferences, rather than storing the entire distribution. These findings support the view that the brain trades off accuracy and computational cost, utilizing structure in sequences of queries to make efficient use of its limited cognitive resources.

CHAPTER 7: Amortized inference strategies give rise to contextual heuristics

The material in this chapter was previously published in Dasgupta et al. (2019b).

The sampling-based accounts discussed so far cannot account for a crucial characteristic of many biases observed in human inference: depending on the domain, they sometimes go in opposite directions. While some studies suggest that people underreact to prior probabilities (base rate neglect), other studies find that people underreact to the likelihood of the data (conservatism). While these have separately been modeled as different heuristics, it is unclear



(a) Memoryless inference **(b)** Amortized inference

Figure 2: Schematics of inference methods. (a) Memoryless inference optimizes the parameters ϕ of the posterior distribution Q_ϕ for each query d. (b) Amortized inference shares parameters across queries, optimizing them such that Q_ϕ is a good approximation *in expectation* over the query distribution.

how these heuristics are learned. Further, the problem of strategy selection remains—it is unclear why and how one heuristic is chosen in certain domains, while a different one is applied in others.

Chapter 7 develops a theory for how heuristic strategies can emerge from the amortization of previous inferences. Amortization is represented schematically in Figure 2. The framework used here is implemented with a recognition model that maps queries d to probability distributions (parameterized by ϕ and defined over the space of hypotheses $\mathcal{H}$).[1] This function is learned from previously computed solutions, thereby 're-using' them. Beyond the direct reuse of old inferences discussed in Chapter 6, learning such a regression function (from d to ϕ) allows us to partially reuse computations for queries not identical, but 'similar' to previous ones. The amortized recognition model provides good posterior estimates *in expectation* over the distribution of queries. This fosters dependence on the underlying statistical structure in the distribution of queries encountered—frequent queries will be better represented than infrequent ones. If the distribution of queries varies across domains, then different domain-specific heuristic strategies (like base rate neglect or conservatism) emerge. Which heuristic arises in which domain depends on how well it captures the structure of the query distribution in that domain—i.e. on how ecologically rational it is.

This chapter shows (with 9 new simulations) that the predictions from an amortized recognition model can reconcile decades of contradicting findings on context-dependent reactions to prior and likelihood (Benjamin, 2018). It can also replicate effects of sample size (Griffin and Tversky, 1992) and experimental design (Koehler, 1996) on these reactions. Further, it tests new predictions from this model by eliciting different reactions to prior and likelihood in the same domain, simply by manipulating the historical query distribution—with a novel experiment as well as a reanalysis of data from Gershman (2015). This framework also explains several related effects including belief bias (Cohen et al., 2017) and similarity weighted reuse (Dasgupta et al., 2018b). Finally, this chapter frames amortization as a regularizer for other noisy inference algorithms (see Zhu et al., 2018, for a similar argument). This framing permits integration of amortized inference strategies (based on variational inference) with the sampling-based algorithms discussed in Chapters 5 and 6 (as the point of initialization of a Markov chain, or as

[1] We actually learn a variational approximation to the true posterior since this can be learned without knowing the true posterior and instead only knowing the joint distribution.

a proposal distribution), providing new unified models for human probabilistic inference.

CHAPTERS 8 – 9: Ecological rationality in machines

Recent years have seen vast improvement in the capabilities of artificial intelligence systems, driven primarily by developments in deep neural networks (LeCun et al., 2015). However, the lack of structure in the representations and decision criteria these systems learn continues to engender skepticism about their interpretability (Doshi-Velez and Kim, 2017), and their promise as a general approach to artificial intelligence (Marcus, 2018; Lake et al., 2018). These two chapters show how ecological rationality provides a new lens for understanding and improving machine learning. These systems implicitly amortize computations, and in doing so have a strong dependence on (potentially spurious) statistical structure in the query distribution. A closer analysis of their learning environments therefore provides insights into their internal representations. In addition, we can manipulate learning environments to engineer new forms of intelligence, like compositional representations and causal reasoning. This also provides insights into the analogous capabilities in humans.

CHAPTER 8: Compositionality in sentence representations

The material in this chapter was previously published in Dasgupta et al. (2018a) and Dasgupta et al. (2019a).

Language exemplifies a hallmark of human intelligence: the ability, in the words of von Humboldt, to "make infinite use of finite means." This ability has been formalized as systematicity (Fodor and Pylyshyn, 1988; Lake et al., 2019): an algebraic capacity to produce new combinations from known components, generalizing knowledge from one context to others. This chapter examines if artificial language representations are similarly systematic. With initial results indicating lack of systematicity, it then explores whether systematicity can be learned in these systems by making augmentations to the training environment, i.e. by making certain forms of systematicity more ecologically rational. Finally, this chapter studies properties of the representations learned with these augmentations, discovering new similarities and differences with human representations of language.

This chapter focuses on sentence representations learned by a machine-learned system for natural language processing (Conneau et al., 2017). By building and analyzing performance on a new diagnostic test dataset, we see that the learned representations are not systematic. Rather, they employ simpler heuristics. Analyses of the training distribution reveals that, due to spurious structural regularities in how the data was generated, these simple heuristics are in fact 'ecologically rational'. Using this approach—of studying the learning environment to gain insight into the representations learned by machine learning systems—I demonstrate new paths towards improved interpretability.

Given that we have control over the learning environment of these systems, we can augment them to alter which heuristics are ecologically rational, analogous to generating adversarial examples (Goodfellow et al., 2014). This chapter shows that if the ecological validity of the discovered heuristics is reduced, the system does learn a more systematic and compositional representation of sentences. Further analyses of these representations reveal parallels to the analogous representations in people. While these systems can learn abstract, systematic rules

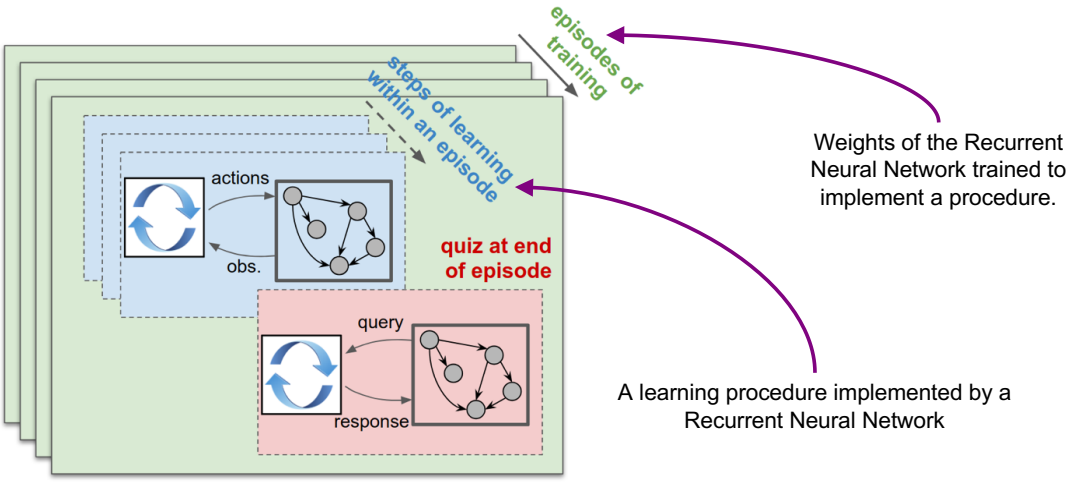


Figure 3: Two loops of learning in meta-learning. The circular arrows represents our learning system, a Recurrent Neural Network (RNN). The directed graph represents the structure of a task. The inner loop of learning (executed by the RNN) optimizes performance in this particular task. The outer loop trains the weights of the network (over related tasks) to implement this inner loop procedure.

and generalize them to new contexts under certain circumstances (similar to human zero-shot reasoning), this generalization has some shortcomings. Notably, these shortcomings are similar to deviations from normativity found in humans (for example belief bias, as studied in Chapter 7). These parallels suggests new ways to understand psychological phenomena in humans. The new metrics presented here (for testing and characterizing the systematicity of language representations) provide inroads toward a clearer picture of what 'human-like' language understanding is, and provide concrete milestones on a path to artificially replicating it.

CHAPTER 9: Learning causal inference from the environment

The material in this chapter was previously published in Dasgupta et al. (2019c).

Traditional approaches to building artificial intelligence focus primarily on engineering new models and architectures that can make more efficient use of computational resources to learn complex concepts and behaviors, on fairly standardized datasets. By considering the role of the environment in shaping inference, we open up a new set of ways to engineer artificially intelligent systems by directly manipulating their training environment. This chapter demonstrates how a simple neural network architecture (trained with trial and error learning from reinforcement) can exhibit causal reasoning and active information seeking behaviors, if we engineer its training distribution.

These experiments are carried out under the meta-learning, or 'learning to learn' framework (Schmidhuber, 1987; Thrun and Pratt, 2012). Here, rather than learning to perform a single task, systems encounter a series of related tasks. Over this experience, they learn commonalities across these related tasks that allow them not only to become better at solving each task at hand, but also to solve previously unobserved tasks from the same distribution with little new experience. A schematic of this framework is presented in Figure 3. Controlling the distribution of tasks the agent encounters therefore provides a way to indirectly control the learning and inference procedures the system encodes.

The absence of causal sensibilities in modern machine learning has been a long standing criticism (Pearl, 1988, 2000). I show that meta-learning agents trained this way can learn strategies that effectively probe, uncover, and leverage the specific kinds of causal structure in their environment to perform causal reasoning in related held-out tasks. They can also select informative interventions, draw causal inferences from observational data, and make counterfactual predictions. This work lays the foundation for causally directed, structured exploration in artificial intelligence, using agents that can perform and causally interpret experiments in their environments, much like human active learning (Nelson, 2005; Montessori and Holmes, 1912). It also suggests exciting new theories for how causal reasoning emerges in humans.

Conclusion

Herb Simon (1955) put forth the challenge facing more realistic theories of 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 takes steps toward exactly that. By taking into account the circumstances under which intelligent behavior manifests, this thesis provides and furthers several new insights. It first presents ecologically rational computational models of human probabilistic inference that leverage environmental structure (via flexible reuse of previous computations) to reduce computational costs. These can therefore explain how we solve very difficult problems with such stark limitations on time, data, and energy. This very adaptation comes at a cost: by adapting to the ecological distribution of queries, we become better at computing good approximate solutions to frequent queries, but worse at answering infrequent ones. This explains a plethora of cognitive biases and deviations from normativity in human probabilistic inference. My models adapt to statistical structure via amortization of inference within structured probabilistic models, providing a path to reconciliation between the historically incompatible statistical and structured approaches to cognition, incorporating their complementary advantages. Further, the lens of ecological rationality provides new insights and inroads into artificial intelligence. This thesis shows that analysis and control of training datasets for machine learning helps us understand black-box systems, providing solutions to the interpretability crisis in modern machine learning. I also demonstrate the emergence of new kinds of intelligent behavior, like causal learning and compositional representations, via manipulation of the training environment. These novel approaches to eliciting such complex behaviors also suggest new theories for how humans acquire and implement them. By building computational theories for the interaction between intelligent systems and their environments, i.e. by developing algorithmic approaches to ecological rationality, this thesis jointly furthers the closely intertwined goals of understanding human intelligence, and building artificial systems that emulate it.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Psychology Press.

Benjamin, D. J. (2018). Errors in probabilistic reasoning and judgment biases. Technical report, National Bureau of Economic Research.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, 50(3):255.

Cohen, A. L., Sidlowski, S., and Staub, A. (2017). Beliefs and Bayesian reasoning. *Psychonomic Bulletin & Review*, 24(3):972–978.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv*.

Dasgupta, I., Guo, D., Gershman, S. J., and Goodman, N. D. (2019a). Analyzing machine-learned representations: A natural language case study. *arXiv preprint arXiv:1909.05885*.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., and Goodman, N. D. (2018a). Evaluating compositionality in sentence embeddings. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Dasgupta, I., Schulz, E., and Gershman, S. J. (2017a). Where do hypotheses come from? *Cognitive Psychology*, 96:1–25.

Dasgupta, I., Schulz, E., Goodman, N., and Gershman, S. (2017b). Amortized hypothesis generation. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Dasgupta, I., Schulz, E., Goodman, N. D., and Gershman, S. J. (2018b). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, 178:67–81.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., and Gershman, S. J. (in press, 2019b). A theory of learning to infer. *Psychological Review*.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., and Kurth-Nelson, Z. (2019c). Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Edwards, W. (1968). Conservatism in human information processing. *Formal Representation of Human Judgment*, 17:51.

Fernbach, P. M., Darlow, A., and Sloman, S. A. (2011). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition*, 119(3):459–467.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71.

Fox, C. R. and Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science*, 44(7):879–895.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS computational biology*, 11(11):e1004567.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278.

Gigerenzer, G. and Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Goldstein, D. G. and Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Griffin, D. and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24:411–435.

Griffiths, T. L. and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17:767–773.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19(1):1–17.

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6:107–118.

Lake, B. M., Linzen, T., and Baroni, M. (2019). Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2018). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Lieder, F. and Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, pages 1–85.

Lieder, F., Griffiths, T. L., and Goodman, N. D. (2013). Burn-in , bias , and the rationality of anchoring. *Advances in Neural Information Processing Systems 25*, 25:1–9.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Montessori, M. and Holmes, H. W. (1912). *The Montessori Method: Scientific Pedagogy as Applied to Child Education in" The Children's Houses"*. Frederick A. Stokes Company.

Nelson, J. D. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 112(4):979.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2):129.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., and Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(3):573–582.

Thrun, S. and Pratt, L. (2012). *Learning to learn*. Springer Science & Business Media.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Vul, E. and Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, 19:645–647.

Windschitl, P. D. and Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1):198.

Zhu, J.-Q., Sanborn, A. N., and Chater, N. (2018). Bayesian inference causes incoherence in human probability judgments. *PsyArXiv*.