Dissertation Advisor: Professor Samuel J. Gershman Ishita Dasgupta

Algorithmic approaches to ecological rationality in humans and machines

Abstract

How do humans reason intelligently in a complex and ever-changing world within limits on energy, data, and time? How can an understanding of this help us build human-like artificial intelligence? Structured Bayesian models provide a normative account of rational behavior. Although computing rational responses via exact Bayesian inference is expensive, empirical findings show that human behavior is often consistent with these rational responses. This seems to indicate that an efficient inference engine underlies human cognition. However, in several notable cases, humans display 'cognitive biases', where their judgments deviate systematically from exact Bayesian inference. How might these contradicting findings be reconciled? This thesis provides such a reconciliation by building on the insight that humans are not general purpose computers: we are instead 'ecologically rational', adapting to structure in our environments to make the best use of our limited computational resources. Chapters 3–4 discuss algorithms for approximating exact Bayesian inference within limitations on computational resources. These reduce the costs of inference by leveraging underlying environmental structure through a process of *amortization* (the adaptive re-use of previous computations). However, amortization can lead to errors when the current query is not representative of past experience. Chapters 5–7 demonstrate that these errors replicate several human cognitive biases. New predictions are tested in several behavioral experiments. Chapters 8–9 demonstrate that amortization also gives rise to ecologically rational behaviors in machine learning, and show how this can be leveraged to artificially engineer new kinds of intelligent behaviors like causal reasoning and compositional language representation. This also provides new insights into how these central tenets of intelligence arise in humans. By taking an algorithmic approach to ecological rationality—i.e. making explicit claims about how it can be implemented at the level of computational processes—this thesis develops new models for human probabilistic inference that can explain both its remarkable successes and seeming failures, as well as suggests new avenues toward machines with human-like intelligence.