Abstract

There are two longstanding traditions in cognitive science -- spanning philosophy of mind to computational cognitive modeling -- of explaining human behavior as arising from ideal rational reasoning or from heuristic algorithms. Attempts to reconcile these views generally treat algorithmic models as approximations of rational ones, constrained by the brain's limited computational resources. However, this perspective is at odds with findings that simple algorithms that completely ignore aspects of prior experience (training data) sometimes yield objective performance superior to rational ones. These seemingly paradoxical less-is-more effects lead us to a new view of algorithmic models, as limits of rational ones with infinitely strong inductive biases (Bayesian priors). We show that influential algorithmic models of decision making, category learning, and reinforcement learning can be derived in this way, yielding new insights into their operation and predictions. We also show how one can identify the inductive biases embodied by existing algorithmic models and define Bayesian relaxations of them, thereby obtaining new and more robust rational models for psychology and machine learning.