

Intrinsic Motivation, Cumulative Learning, and Computational Reinforcement Learning

Andrew Barto

Computer Science Dpt, University of Massachusetts Amherst, USA

Motivation refers to processes that influence the arousal, strength, and direction of behavior. Psychologists distinguish between extrinsic motivation, which means doing something because of some specific rewarding outcome, and intrinsic motivation, which refers to doing something because it is inherently enjoyable. Intrinsic motivation leads organisms to engage in exploration, play, and other behavior driven by curiosity in the absence of externally-supplied rewards. Intrinsically motivated learning has long been viewed as essential for the cumulative development of an agent's competence in interacting with the world. In this talk, I review some of the research directed toward developing intrinsically motivated learning systems, which is not at all a new idea though it is receiving increasing attention. I focus in particular on how to design intrinsically motivated reinforcement learning systems. I discuss five themes that stand out as being important: 1) the distinction between a reinforcement learning agent and its environment at the base of the computational reinforcement learning framework has to be looked at in the right way; 2) internal state components that influence intrinsic reward include a robot's memories, beliefs, value function, and policy in addition to vegetative features like battery and dust bin levels; 3) a guiding principle is that the learning and behavior generation processes "don't care" if the reward signals are intrinsic or extrinsic; the same processes can be used for both; 4) the dividends paid by intrinsically motivated reinforcement learning accrue over multiple specific tasks faced over extended periods of time; and 5) intrinsically motivated reinforcement learning is a good way to produce behavioral modularity that is essential for cumulative learning.