



# Research Statement

Cameron Allen

 camallen.net  
 csal@brown.edu

My goal is to both understand the computations that enable intelligence, and harness those computations to produce safe and beneficial general-purpose AI. Generally intelligent agents must learn and plan in complex environments, with sensors and actuators that support various behaviors and tasks. This complexity hinders decision making, necessitating methods that abstract away such intricacies. My research is mainly focused on managing the complexity through practical algorithms that learn provably useful abstractions. This problem is challenging, because the usefulness of such abstractions depends on both the specific use-case of the AI agent and the need to interface with the human end-user. In my dissertation, I investigated several ways to build agents that automatically learn useful decision-making abstractions for both reinforcement learning and planning.

In reinforcement learning, I developed a method that learns abstract representations to deal with noisy, high-dimensional observations. The method overcomes longstanding challenges in representation learning with a practical training objective that provably retains sufficient information for decision making. I showed that the learned representations are human-interpretable and lead to state-of-the-art performance across a wide range of applications. The work was published at NeurIPS 2021, and I am presently extending it to handle two new cases: one where even the agent's rich observations are insufficient and must be augmented with memory, and another where the representations must support learning a compositional model of the world.

In planning, I introduced two novel skill-based abstraction methods that can improve planning efficiency by orders of magnitude. The first project, published at IJCAI 2021, constructed a library of skills with focused, interpretable effects and minimal side-effects, which allowed the planning algorithm to make substantially better use of a simple but computationally efficient search heuristic. The next project, currently under review, considers libraries of skills so large and so general-purpose that they may render planning intractable for any specific problem. My colleagues and I designed an algorithm that carefully restricts the set of skills under consideration to only those relevant for optimally solving a particular task, and we showed empirically that this can reduce planning time by a factor of 75 and reduce search space size by a factor of 280.

Looking forward, I want to understand how theoretically-principled abstraction methods can be used to improve both our ability to build—and our understanding of—complex AI systems. For example, in the context of AI alignment, I am interested to see if abstraction techniques can help make an agent's decisions more interpretable by human auditors, and bring its internal objectives into better agreement with external human-provided feedback. I am also interested in what kinds of abstract representations best support modeling an agent's uncertainty about human preferences. Related to these questions, I am curious to what extent an agent's representations can be made to resemble human representations, in order to better facilitate collaborative problem solving.

I am well suited to tackle the challenges related to building safe and human-compatible AI. My experience designing and building agents that learn provably beneficial abstractions has given me a strong foundation not only in theory and mathematics, but also in engineering software systems and executing empirical evaluations. A postdoctoral research fellowship at CHAI would draw on these skills and allow me to expand my research portfolio as I prepare for a future faculty position.