# Abstract

Artificial agents are technological artifacts that perform tasks for people. They sense the world and relate perceptions to their tasks in order to identify and then apply an appropriate response. In general, we want to build agents that perform increasingly important and complex functions outside of human supervision. However, our ability to enhance agent autonomy depends upon advances in the state of the art of agent design, while our willingness to deploy such systems demands new methods of validation. Human trust limits agent autonomy.

This thesis defines a novel architecture for artificial agents that increases their autonomy while enhancing our trust. Value-driven agents are unique because they act to maximize an internal measure of reward. This design allows increased autonomy by motivating agent behavior from a sense of what is important instead of a predetermined task. The reward function also provides a guide for learning outside of human supervision. We enhance trust by supplying a theoretical guarantee: a well-aligned value-driven agent will maximize human utility as a consequence of learning to maximize its own reward. This is the first instance of a guarantee that spans the reference frames of artifacts and humans.

A value-driven agent consists of a reward function and a set of skills encoded in a reactive language that embeds a learning algorithm. The language (Icarus) and the learning algorithm (SHARSHA) are new technology. Icarus adds methods for state, operator, and goal abstraction to reactive designs, and supports value-based choice among the options within skills. SHARSHA is a hierarchical reinforcement learning algorithm mated to Icarus plans. We prove its convergence properties. The combination of Icarus and SHARSHA contributes a novel and general method for embedding domain knowledge in reinforcement learning problems. We obtain a formal bridge between agents and users by combining the SHARSHA proof with a discussion of value alignment. If an Icarus agent can learn to maximize its own reward and that reward is aligned with user concerns, the value-driven agent will resolve the best strategy within its ability to maximize user utility. We call this a ‘Be all you can be’ guarantee. It validates

agent behavior in advance of learning, and increases our willingness to deploy highly autonomous systems.

We conduct two experiments in a simulated vehicle control domain to demonstrate the benefit of the value-driven architecture. The first examines the effect of encoding domain knowledge in reinforcement learning problems. We conclude that additional distinctions about state improve performance but decrease learning rate, while additional plan structure can increase both learning rate and performance. Plan structure also decreased plan size by three orders of magnitude relative to the expected formulation of our test problem. This suggests a qualitative change in the scope and efficacy of feasible learning applications. The second experiment examines the benefit of the value-driven architecture for agent design. We show that different reward functions can generate qualitatively different behavior over the same set of skills. This provides evidence for the feasibility of a novel design method: we can develop one fixed skill base for an application area, and customize individual agents via programming by reward.