Multi-Agent Cooperation via Intrinsic Motivation

Blake Elias

March 2, 2021

1 Utility as Surprise

Standard reinforcement learning puts forth an agent whose actions depend on two objects: its model of the environment along with its utility function. Rational choice theory predicts that the agent will take those actions which it predicts will result in the highest-utility pay-off. However, the question of where this utility function comes from biologically, as well as how we are supposed to model it when studying social systems, has remained largely un-addressed by social science literature.

To solve this problem, proposals have been made for ways to model intelligent agents, including human agents, without the use of a separate utility function. Instead, agents can be seen as optimizing for "expected surprise": being on a mission to create the best possible model of the environment, so as to have the highest degree of control and therefore the highest degree of predictability. Surprise can be formulated as the (in)ability for an agent to predict the consequences of its own actions. [**?**] This provides an entropic measure which can be used in place of, or in addition to, the standard utility function.

Such utility functions have also been provided as "intrinsic rewards" to digital reinforcement learning agents [**?**], with the result of helping agents learn advanced game behaviors which were otherwise un-learnable due to the "sparse rewards" problem in certain game contexts (e.g. getting only a single binary reward at the end of a level, based on survival or death).

2 Multi-Agent Utility

In multi-agent contexts, there have been even greater challenges than the standard single-agent reinforcement learning setup. In particular, there has been difficulty in formulating clear frameworks for dealing with cooperative or competitive relationships between agents. From the perspective of any one agent, the environment is non-stationary (i.e. not describable by a Markov Decision Process) due to the presence of other agents who may take unpredictable actions. Indeed, the other agents can become the most important—and least predictable—part of the environment. Similarly, from the perspective of a central planner, there is no single clear way to frame the optimization problem.

Existing approaches include maximizing total reward (i.e. summing all agents' reward functions, forming a "social welfare" metric of sorts), or treating the problem as multi-objective optimization (i.e. each agent's reward as a separate objective) for which a Pareto front—but not a single optimal—can be described.

One key source of difficulty in such multi-agent scenarios is the possibility of agents having different, and unknown, reward functions. This makes it difficult to (a) predict the actions of other agents in a given scenario, and (b) determine which agents' rewards should be prioritized in non-cooperative scenarios.

2.1 Shared Goals to the Rescue

Could the entropic, "surprise-minimizing" view of utility lend new insight in multi-agent scenarios? What simplification can be had by treating agents' utility functions not as arbitrary and unknown, but instead as having a particular functional form that is known *a priori*? What advantage might we gain with respect to challenges (a) and (b) above?

Firstly, let me note that, despite the assumption that the agents' utility functions take similar form, this does not mean that they are identical. This is because the agents each take on a different structure, both at the morphological and cognitive levels. Due to morphological differences, the capabilities of perception and action in a given situation will be different for each agent. So too, therefore, will the agents' level of expected surprise be different, even when the same world state is presented. (E.g. A certain temperature increase may be imperceptible to a turtle, yet very shocking and debilitating to a human. Conversely, weather changes that humans are able to easily adapt to via clothing and shelter may pose great threat for a lion who does not possess these same abilities.) Furthermore, the agents' diverse experiences have led them to different states of knowledge. Thus, even two morphologically similar humans, such as twins, may experience different levels of "surprise" in the same world state when a given action is taken, due to their differing cognitive states.

Thus, while each agent still has a unique utility function, these unique utility functions are related to one another, are non-static, and can be aligned via changes in knowledge. Indeed, the primary cause for differing utility functions among agents with similar morphologies, is due to a difference in cognitive states. One means of aligning incentives between agents, therefore, is to communicate and share knowledge.

With this understanding, we can envision partial solutions to the two problems described above: (a) predicting the actions of other agents, and (b) determining how to prioritize their different reward functions. We cannot fully solve either of these challenges fully, as we are limited by how directly we can measure an agent's morphological or cognitive state. But to the extent that we can measure or infer these states, we can form partial knowledge of the agent's current utility function.

Regarding challenge (a) of predicting action: even knowing an agent's utility function perfectly still does not enable perfect prediction of the agent's behavior. Firstly, this is because the agent may act sub-optimally in some unknown way

while optimizing for their own utility function. Secondly, there can be multiple ways to optimize the same utility function. Thus, while knowing something about an agent's utility function does not fully determine its action, it may put some bound on the space of what actions it might reasonably take.

For challenge (b), knowing that the agents' utility functions have some shared structure might lend a clearer definition of what it means to optimize for the group outcome. The utility function of the entire multi-agent system might be represented as the summation of the utility functions of all agents.

2.2 Education for Cooperation

Initially, the "sum of utilities" can yield the same conflicting-incentives problem we started with, where what is best for the collective may not be best for any given individual. Yet we are armed with a new tool: the understanding that agents' incentives will *change* once their cognitive state—i.e. their model of the world—changes. Thus, a group of morphologically similar agents, starting with different knowledge, ought to converge on nearly equivalent utility functions, assuming all agents employ a learning algorithm that are guaranteed to converge on a consistent model of the world. To facilitate this process, we can introduce more actions into the action space: namely, a set of *teaching* actions that allow an agent to communicate part of its knowledge to another agent.

There has been much work in the multi-agent literature on emergent communication strategies when channels are made available between the agents. However, I propose here that these channels might be more effectively used if the agents are endowed with a model of (1) *how the other agents learn* (both from the world itself and from communication passing), and (2) *how agents form utility function* based on their cognitive state. The problem of an agent achieving an optimal outcome for itself can thus become a model-based planning problem, involving two types of actions: those that directly provide some individual utility, and those that teach and learn from other agents so as to better align knowledge (and therefore, incentives). With better-aligned incentives, the agents will find themselves better able to engage in collective action, increasing both individual and group reward.

2.3 Pre-Conditions for Cooperation

The link from knowledge-sharing to incentive-alignment assumes two things shared in common between the agents: (1) similar morphology (perception and action ability), and (2) mutually understood, if not identical, learning algorithms (to ensure that teaching can indeed have its desired effect). If either of these conditions is broken, then an agent may find itself *worse off* by being part of a given group, being "dragged down" by the group's different physical needs, different learning/communication styles, or both.

3 Contributions

I anticipate that the knowledge of how beliefs are formed (the subject of learning theory), how those beliefs give rise to desires (the "surprise-minimizing" theory of utility), and a theory of how desires plus beliefs give rise to action (the subject of decision theory), will soon provide a rich toolbox for reasoning about cooperation in multi-agent systems.

The insights here thus provide a "big-picture" overview of how learning theory, decision theory, and entropic models of utility interlock to provide the foundation for a cooperative theory.

4 Future Work

Many questions are left open, largely in the following areas:

- How differences in agent morphology affect the utility agents have.
- How agents with similar learning algorithms can learn to teach one another, and therefore, cooperate.
- How differences in agents' learning algorithms can affect their compatibility for existing within a collective structure.
- How, or when, an agent should ever decide to leave its current group. To what extent does this depend on its physical abilities (i.e. one cell in a multicellular organism cannot simply "walk away" from the other cells once it is part of a connected tissue, while humans in a society can to some extent leave their present conditions more easily)? If one agent can only *partially* depart from the other agents, but will nonetheless be affected by those agents' actions (e.g. a human still being affected to some extent by climate change no matter where on Earth it may travel), how does it arrive at an optimal decision as to the proper level of cooperation?