White paper: Value alignment in autonomous systems

Stuart Russell, Professor of Computer Science, UC Berkeley

Summary

For an autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans. I propose to investigate a formal definition of this approach as *cooperative inverse reinforcement learning*, to develop algorithms for solving such problems, and to investigate their behavior in a variety of settings.

Problem statement

One way to subdivide the problem of designing autonomous intelligent systems is first to build a generic decision-making capability and then to supply the necessary elements of the decision problem the system should address: the transition and sensor models and the reward or utility function.

Whereas the transition and sensor models are in a sense task-independent and can often be constructed from the basic physics of the problem or learned from abundant empirical data, the reward or utility functions *constitute* the task – they are the agent's only source of information about what it is supposed to do.

There are many examples of simple tasks – such as video games – where defining an appropriate reward or utility function is easy. For more complex tasks in unstructured environments, particularly those involving humans, defining rewards (and hence defining optimal behavior) is much more difficult. For example, an autonomous vehicle that drives on ordinary streets must understand the tradeoffs involved among many aspects of its environment, including travel time, fuel efficiency, legality, predictability by other drivers, passenger comfort, passenger safety, safety of other drivers, pedestrians, pets, wild animals (large and small), inanimate objects (large or small, soft or hard, valuable or worthless), etc. Similar considerations apply in military contexts, where the stakes are arguably even higher.

As several authors have pointed out (Omohundro, 2008; Yudkowsky, 2011; Bostrom, 2014), a mismatch between the defined reward function of an intelligent system and broadly shared human values can result in extreme violation of those values. Mathematically speaking, this is a typical consequence of optimizing when the stated objective function depends on only a subset of the available variables: the other variables will often be set to extremal points. Without a solution to this problem, building advanced intelligent agents may expose humanity to catastrophic risk.

Solution approaches

I have previously proposed *inverse reinforcement learning* or IRL (Russell, 1998; Ng & Russell, 2000) as the problem of acquiring a reward function from observation of another agent: assuming that agent is behaving optimally, what reward function best explains its behavior? The approach has been applied to many domains with some success, and there is a steady stream of theoretical results and novel formulations, but it is not quite the right answer to the general problem. The reason is that we do not necessarily want a robot[1] to adopt the *same* reward system as the humans[1] in its environment. For example, if a human appears to be trying to make a cup of coffee, it is reasonable to assume the human wants coffee; but we don't want the robot to want coffee.

The difficulty seems to be that IRL envisages a single agent (initially the expert human) in the environment, and the robot is training to *be* that agent. (Work on multiagent IRL by Natarajan *et al*. (2010) does not avoid this problem.) Instead, the robot should be learning about the human's reward function in order to maximally helpful *for the human*. One might call this the *cooperative inverse reinforcement learning* or CIRL problem. The following sections describe two possible directions for solving it.

Single-agent formulation

One obvious solution is simply to solve the IRL problem as before, but to use the learned reward function indexically – i.e., the robot is rewarded when the state is rewarding *to the human*. For example, when the human finally drinks a cup of coffee, perhaps with the robot's assistance, the robot obtains an equivalent reward. The robot's model of the environment includes both the human and itself. One could add to the robot's reward function some elements reflecting self-preservation, but in principle those elements are unnecessary because they are already reflected in the value the human places on the robot's existence.

This approach is similar to the formulation of the *decision-theoretic assistance* problem studied by Fern *et al*. (2014). In that work, the human is assumed to have a discrete goal (e.g., a destination) rather than a reward or utility function, and the problem is formulated as a *hidden-goal MDP* or HGMDP, a class of POMDPs in which only the human goal variable is unobserved. Surprisingly, no connection is made to the literature on IRL. It should be possible to construct a formal mapping between HGMDPs and CIRL problems, allowing the more flexible formulations of IRL and the more scalable techniques of machine learning to be applied to assistance problems.[2]

[1] We use "robot" henceforth to stand for any agent that is learning to help some other agent or agents, which we term the "human".

[2] This work might be carried out jointly with Fern and Tadepalli, with whom I have collaborated in the past.

Game theoretic formulation

Fern *et al*.'s work on decision-theoretic assistance is a significant step forward but its formulation as a single-agent decision problem means that the human must be modeled as a stochastic process rather than an agent operating in a multiagent context. This leads to some awkward issues including turn-taking and the assumption that "the [human] is obliged to accept the helper action if it is helpful for its goal and receives a reward bonus (or cost reduction) by doing so."

In order to accommodate the fact that both robot and human will be acting in a multiagent setting, I propose to investigate a cooperative game-theoretic formulation of the CIRL problem, in which the robot must learn a policy – possibly mediated by the acquisition of a reward function and an understanding of the reward function of the human – such that some or all of the Nash equilibria of the cooperative game maximize the human's payoff. (Presumably the traditional ambiguity caused by the presence of multiple equilibria can be reduced or eliminated if the human can reliably assume the robot is acting in the human's interest.) I am particularly interested in extending the Bayesian IRL formulation (Ramachandran & Amir, 2007), which provides a natural basis for formulating optimal exploratory policies whereby the robot actively attempts to discover the human's preferences.

Research activities

In addition to constructing the necessary formal frameworks and analyzing their relationships to each other and to the decision-theoretic assistance model, I will pursue versions of the standard kinds of theoretical results for IRL: algorithms that provably converge to optimal behavior in the limit and PAC-style regret bounds relative to a robot that already knows the human reward function. The algorithms and theorems should also be robust to imperfect human behavior.

I will also investigate the qualitative nature of the rational learning process, particularly in the Bayesian setting with very broad priors over the human reward function and a high degree of risk aversion built in. In such a setting, the rational behavior for the robot should be to find out as much as possible about human preferences while minimizing intervention (in order to avoid inadvertently causing a highly negative outcome for the human). As the robot becomes more confident in its assessment of human preferences, it can start to take actions that it is sure are helpful. These qualitative properties can be investigated initially in simple, artificial experiments. Such experiments can also help in formulating new theoretical results, such as bounding the probability of any outcome that deviates significantly in value from the human's actual preferred outcome.

Another advantage of the Bayesian setting is that, thanks to hierarchical priors and the availability of communication networks, it should be possible to accommodate simultaneous learning by multiple robots interacting with multiple humans in many different scenarios and locations. Humans can be assumed to share a good deal of the core content of their reward functions, even if they differ in many details. Moreover, one may expect a "cleanup effect" (cf. Sammut *et al*., 1992) whereby persistent deviations from rationality in one individual – which can lead to learning an erroneous reward function – can be outweighed by experience gleaned from many other humans.

The long-term goal of this research is to derive theoretically well-founded mechanisms that allow the intelligence of AI systems to be increased, perhaps beyond human levels, with *no risk* to humanity. The value alignment problem urgently requires solutions, however, even for AI systems that are considerably less intelligent than humans, if such systems are to participate safely in unstructured human environments. Thus, there are substantial economic incentives for solving the CIRL problem. I hope to develop an understanding of how robust the solutions for "subintelligent" systems are when applied to systems that are approaching "superintelligence."

References

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Fern, A., Natarajan, S., Judah, K., and Tadepalli, P. (2014). A Decision-theoretic model of assistance. *JAIR*, 50.

Natarajan, S., Kunapuli, G., Judah, K., Tadepalli, P., Kersting, K. and Shavlik, J. (2010). Multi-Agent Inverse Reinforcement Learning. In *Proc. ICMLA-10*.

Ng, A. Y. and Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning.*

Omohundro, S. (2008). The basic AI drives. In *AGI-08 Workshop on the Sociocultural, Ethical and Futurological Implications of Artificial Intelligence*.

Ramachandran, D. and Amir, E. (2007). Bayesian inverse reinforcement learning. In *Proc. IJCAI-07*.

Russell, S. J. (1998). Learning agents for uncertain environments. In *Proc. COLT-98*.

Sammut, C., Hurst, S., Kedzier, D., & Michie, D. (1992). Learning to fly. In *Proc. ICML-02*.

Yudkowsky, E. (2011). Complex value systems are required to realize valuable futures. Technical report, Machine Intelligence Research Institute.