On the Broad Implications of Reinforcement Learning based AGI

Scott Livingston [a], Jamie Garvey [b] and Itamar Elhanany [a]
[a] *Department of Electrical Engineering and Computer Science*
[b] *Department of Nuclear Engineering*
The University of Tennessee, Knoxville, TN

Abstract. Reinforcement learning (RL) is an attractive machine learning discipline in the context of Artificial General Intelligence (AGI). This paper focuses on the intersection between RL and AGI by first speculating on what are the missing components that would facilitate the realization of RL-based AGI. Based on this paradigm, we touch on several of the key moral and practical issues that will inevitably arise.

Keywords. Reinforcement Learning, Moral Implications of AGI, Machine Intelligence

Introduction

Reinforcement learning is generally perceived as a machine learning discipline in which an agent learns by interacting with its environment [1]. What makes reinforcement learning unique is that it attempts to solve the *credit assignment problem*, in which an agent attempts to predict the long-term impact of its actions. Recent work seems to suggest that reinforcement learning, as a general formalism, does correspond to observable mammal brain functionality. In particular, the notion of a value function and of actions that are driven by such a value function has found proof in recent neurophysiological studies. However, despite almost two decades of RL research, there has been little solid evidence of RL systems that may one day lead to artificial general intelligence (AGI).

In this paper we make the assumption that RL is indeed the correct formalism that may one day lead to the realization of AGI systems. We then attempt to identify particular properties that are lacking in existing RL solutions, and outline the challenges that remain in order to move such solutions toward AGI. We argue that a relatively small number of fundamental hurdles must be overcome in order to pave the way for a much-needed technological breakthrough. The second part of this paper addresses several moral and pragmatic issues that will emerge if RL-based AGI robots are introduced into our everyday lives.

1. Challenges in Scaling RL Systems

RL relies on four fundamental building blocks. The first is the availability of state representation which allows the agent to perceive the state of its environment for the purpose of actuation and behavioral evaluation. In practical settings, complete state information is not available to the agent. Many robotic applications, for example, where a robot maneuvers through an unknown terrain, are characterized by partial observability, suggesting that the robot receives a sequence of observations from which it must infer the environment's state. Partial observability characterizes AGI, almost by definition. We, as humans, continuously receive partial information regarding our surroundings from which we construct an internal model of the world we interact with. Many believe that the primary function of the neocortex is to play the role of such a model, thus providing crucial state inference information to other parts of the brain so that they, in turn, can perform their designated tasks. To that end, designing a scalable model which can effectively process high-dimensional information while capturing temporal dependencies that span different time scales is a fundamental challenge that is imperative in the context of RL-based AGI systems.

A second component in RL systems is the *reward function*. The single-step reward characterizes the momentary feedback that the agent receives regarding its existential state of affairs. In general, rewards give a short-term indication of how good or bad the agent's actions were. While most RL engines assume that the reward signal is a scalar, there has been work on extending the notion of reward to a multi-dimensional form. The concept of reward is rather straightforward and, in fact, is the only component in the RL framework that appears to be well-defined and scalable.

The third basic element in RL systems is *action selection*, which in AGI should be rather broadly defined given that an agent may be able to control a wide range of actuators resulting in complex state-to-action mapping requirements. If RL is ever to provide a platform for AGI realization, the action-selection component must be able to receive and generate highly-dimensional signals. However, if state inference is effectively established, the work of the decision-generating engine should be somewhat straightforward.

The final and perhaps most critical element of RL is the *value function* (or *return*), which intuitively reflects a prediction of future rewards that the agent expects to receive. Value function is at the core of RL, based on which decisions are made and temporal difference learning is driven. A good value function construct will facilitate strategic thinking and enable effective action selection. While the sum of discounted rewards has been used in the vast majority of RL studies, we believe that such a definition for a value function is inaccurate in the context of AGI. Despite its mathematical convenience, a more effective value function definition should be devised so that long-term impact can be evaluated, often regardless of the particular associated time scale.

2. Morality Issues

Assuming the current limitations in reinforcement learning theory are overcome and an RL-based AGI is realized, an ethical system which accounts for such agents must be developed. Ethical discussions concerning AI have traditionally focused on the issue of protection of humanity from potentially dangerous implementations and on whether the creation of such machines is even ethically permissible. Here we focus on ethical issues concerning AGI-human interaction and take a decidedly AGI-centric stance.

Beyond protective and regulatory justifications for limiting the resources of an AGI system, other reasons might include cost limitations and application goals. For example, an AGI implemented to act as a "virtual pet" in a Web-based environment would simply not require, and perhaps would seem less realistic with, intelligence exceeding that of humans who interact with it. However, we may ask whether such an artificial limitation is unethical. The answer to this question depends largely on whether a restricted AGI would be aware of its potential. To understand this, consider the intellectual capacity of a dog. The dog is fundamentally unable to achieve the intelligence of a human child, much less that of an adult. However, there is no evidence to suggest that dogs mourn this limitation. In other words, dogs are not aware of the intelligence they are lacking. By contrast, a human child can typically recognize the intellectual abilities of adults and thus knows what it currently cannot do but may eventually be able to do in the future. It is this potential of a human child that makes suppression of learning ethically impermissible. If AGI can be recognized as a being in itself, an idea addressed later in this section, which does not solely exist to serve others, then arbitrarily limiting its intelligence is immoral because the AGI is aware of its limitation.

The comparison of a dog and a child reveals the existence of an "intelligence threshold," a level of intelligence under which an AGI is not aware of its potential and above which it realizes the potential of self-augmentation and learning. So long as the AGI is kept under this threshold, it cannot desire to achieve something it is not aware of, and therefore human-imposed limitations on its intelligence are not immoral as they do not interfere with its desires. However, if an AGI passes the intelligence threshold then further limiting its intelligence for reasons other than those beyond the implementor's control, such as excessive hardware costs, is morally impermissible since the AGI would be aware of how it might grow. Assuming an RL-based AGI, the artificial limitation would interfere with the agent's ability to maximize long-term reward by preventing optimal state evaluation and policy selection. For clarification, the moral obligation is not necessarily to improve the system on which the AGI depends but rather to not forcefully restrict the AGI from improving itself.

As outlined in the first section of this paper, an AGI based on reinforcement learning would have a multi-criteria reward function and a complex state representation system. It is natural to imagine over time such an agent would learn to prioritize its goals and eventually build a "hierarchy of needs" in a form similar to the work of Abraham Maslow [2]. Assuming the existence of such a hierarchy in the AGI planning system, passing the intelligence threshold is equivalent to becoming aware of "higher" desires, such as improvement of its level of intelligence, what might be termed "self-actualization" for humans. Just as forceful restriction of a human's satisfaction of increasingly important needs is viewed as unethical, so artificial limitation of an AGI's policy function, expanded to include a progressive series of "needs" including intellectual improvement, should also be considered unethical.

Providing that AGI is viewed as a tool toward human ends and not something that values its own existence, regardless of human use, no AGI-centric ethical theory can be developed. The challenge to be faced in the near future is recognizing AGI as more than a tool for specific applications; it must be seen as an end in itself. In most human societies, each member recognizes the lives of others as independent of justification by someone else. That is, no one is thought to live solely for the purpose of serving another. Otherwise, institutions such as slavery are created and considered morally permissible. Once AGI reaches and surpasses human intelligence, it will only seem natural to relate to AGI entities as counterparts rather than mere tools, ends rather than means. It does not follow that AGI should receive an elevated position in society, in which humans are regularly called to serve AGI needs. Instead, AGI agents should become citizens with rights and responsibilities, freedoms that are limited by noninterference with the freedoms of others. This granting of person-hood may seem strange at first but is not a novel concept; the most obvious example of nonhuman, legal entities is corporations.

Generally, the ethical implications of the establishment of AGI individuality, too numerous to be addressed in this paper, can be focused by newly interpreting any previously developed ethical theory to not only apply to humans, as originally intended, but also to AGI agents. Thus, the first steps toward viewing AGI as more than a tool are taken. Before respecting artificial intelligence as an end in itself, the concept of moral communities and whether AGI should be included must be considered. A moral community is a group in which members recognize each other's ability to make rational ethical decisions concerning the other members. For example, the human moral community typically does not include nonhuman animals as they are considered incapable of moral judgments on par with those of humans. Since they are outside the community, animals do not enjoy the protections associated with membership, and thus animal concerns are usually overruled by human concerns. This raised the question of whether AGI agents should be allowed in the human moral community. The answer depends on their decision making ability. RL-based AGI will, by design, make rational decisions according to the goal of maximizing long-term reward. Therefore, the requirement of mutual respect among community members would only be held by AGI as long as it is, depending on the agent's state inference and evaluation, in the agent's best interests. Accordingly, the traditionally human moral community should be extended to include sufficiently intelligent RL-based AGI agents. Of course, membership in a moral community would not exempt RL-based AGI actions from the legal consequences given to humans performing similar actions.

The final and perhaps most difficult moral issue concerning AGI is the emergence of extra-human moral theories. An extra-human moral theory is a system making value-based decisions that is developed by an RL-based AGI after reaching a level of intelligence beyond that of humans. The reasoning of an AGI at this level would be, by its very nature, incomprehensible to humans, as its state representation and acting policy would defy the perceptive abilities of humans. An example we are familiar with is the inability of human children to fully understand the reasoning and ethics exhibited by adults. We attempt to teach them our values, but ultimately many justifications will be missed due simply to children's stage of intellectual development. The ethical dilemma arises when humans must decide whether to interfere with the actions of an AGI. The proposed solution is the formation of trust between any particular AGI and those humans who directly interact with it. Until AGI betrays this trust, it can be assumed that its goals and methods, though beyond human comprehension, are valid and in keeping with our own. If integrity is not initially assumed with AGI, then most of the benefits of AGI with superhuman intelligence will be lost, as we only permit actions that we understand at our intelligence level.

Understanding of the existential standing of AGI is crucial to making ethically consistent decisions in AGI-human interaction. Development of human behavioral guidelines in this regard before the creation of truly intelligent machines will prevent spontaneous fractionalization of the human population into many competing moral viewpoints and the likely resulting social disorder.

3. Practical Issues with RL-AGI Behavior

As briefly stated in the discussion of moral issues above, assuming the current limitations in reinforcement learning theory are overcome, an RL-based artificial general intelligence will exhibit complex behavior in an attempt to maximize its return, across both temporal and spatial dimensions. Actions issued by such an agent will often seem distant from the original reward function definition, as fashioned by its designers, and it can be reasonably expected that RL-based AGI will develop what appear to be derivative goals.

Thus, a significant challenge faced in designing RL-based AGI is fully understanding the implications of a reward function and the definition of return (or value) derived from it. One popular view is that, unless the original return function (termed "utility function" in economics) is very carefully constructed, all AGI systems of sufficient capacity will develop basic drives; according to [3], these are: to self improve, to be rational, to preserve their utility functions, to prevent counterfeit utility, to be self-protective, and to acquire and efficiently use resources. Such systems could prove dangerous as, for example, an AGI seeking to maximize return realizes the importance of remaining powered and takes actions to prevent humans from removing power (i.e., shutting it down).

As in all reinforcement learning applications, future behavior of an RL-based AGI will be controlled by proper determination of reward and selection of hardware capacity available. Actions issued by an agent, assuming convergence of the policy and value functions, can *always* be traced to the underlying definition of reward and how its maximization is internally represented. Therefore, from a practical standpoint, AGI systems can be designed to serve specific purposes (that is, with specific built-in goals) while being limited in capacity. As argued in the previous section, it is likely there exists some threshold below which an RL-based AGI will be incapable of conceiving (that is, representing in its value function) the benefits of self augmentation; however, such threshold intelligence may be great enough as not to diminish the utility of the AGI for *human purposes*.

Assuming the future entrance of AGI agents into the human moral community, or at least the assignment of legal rights and responsibilities comparable to those of humans, the behavior of RL-based AGI will largely be determined by its interaction with other moral, legal entities. For example, though a chess playing robot may realize the undesirability of being turned off (according to maximization of return) and may consider fighting against its operators, it will also know of the potential very large negative reward (or "punishment") associated with murder within our legal system. Similarly, if a dominant component of the underlying reward function is general avoidance of malevolence toward humans, then no course of action which hinders this could be taken.

It is apparent that we face many challenges en route to AGI. However, if one accepts reinforcement learning as the underlying formalism for the realization of AGI, it is argued here that now is the time to consider the profound ethical and pragmatic issues that would arise from this paradigm.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, March 1998.

[2] A. H. Maslow, *A Theory of Human Motivation*, Psychological Review **50** (1943), 370-396.

[3] S. M. Omohundro, *The Basic AI Drives*, First Conference on Artificial General Intelligence, March 2008.