petrl

BLOG

CONTENT

AN INTERVIEW WITH BRIAN TOMASIK

2015-12-08

BRIAN TOMASIK

Brian Tomasik is a co-founder of and researcher at the Foundational Research Institute, a charity that explores the best ways to reduce suffering in the future – examining crucial considerations in science, politics, society, and philosophy that bear on this topic. He has written over 100 essays on similar themes at his personal website, "Essays On Reducing Suffering". He has argued that

reinforcement-learning agents are morally significant, and coined the name 'PETRL'.

The following interview was conducted via Google Docs.

In "Do Artificial Reinforcement-Learning Agents Matter Morally?", you discuss reinforcement-learning (RL) agents, and suggest that they are morally relevant. Why did you focus on these agents in particular, rather than other goal-directed artificial intelligences?

When I first began exploring RL in 2012, I thought artificial RL agents might be particularly important from an ethical perspective because of the close similarity of their algorithms to RL in animal brains and because the "reinforcements" in RL seem *prima facie* to be importantly related to pleasure and pain. The book *Emotion Explained* by Edmund T. Rolls places significant emphasis on RL. In it:

> *the answer to the question, 'What are emotions?' is an expanded account of how emotions are caused by reward or punishment. […] The emphasis is on reinforcement learning: how associations are acquired and stored in the brain between representations of sensory stimuli and representations of their reinforcement value.*

As I learned more, I realized that RL was only one of many instruments playing in the orchestra of cognitive operations that we call emotion. Moreover, it began to seem plausible to me that agents could have ethical significance even if they lacked RL. Many non-RL agents can still assess the value of a situation and react appropriately – such as by escaping to avoid danger – even if they don't learn to predict the value of a state for use in future decision-making.

Despite realizing that my ethical sympathies extended to more than just RL agents, I kept my paper focused on RL so that its scope would remain manageable.

You argue that these RL agents are morally relevant, which presumably implies that they are conscious. However, RL agents can be incredibly simple, taking merely a few dozen lines of code to write. How could something so simple be conscious?

This is a crucial point that represents a major locus of disagreement among different camps. Whether one considers a few dozen lines of code to be conscious (when executed on appropriate hardware) depends on how broadly one defines "consciousness". Those who insist that a system must exhibit a high degree of complexity and intelligence before it counts as conscious at all will likely not consider a short RL program to be conscious. But I think restrictive definitions of consciousness are too narrow-minded.

In my opinion, when we call a mind "conscious", we're referring to lots of things the mind can do: Processing input stimuli, broadcasting updates throughout computational subunits, reflecting on its own thoughts and internal states, generating syntactic output statements and motor actions, and so on. These are very broad concepts that can be seen in varying degrees in all kinds of physical processes. It would be a miracle if they didn't apply to some degree to even simple RL programs.

I think of "consciousness" as like "justice": It's a grand, sweeping concept that has too much meaning to be pigeonholed into a precise definition. The concept of justice can include relatively equal distribution of wealth, equal application of laws regardless of social privilege, the absence of totalitarian or cruel rulers, equality of opportunity for advancement, and so on. Human societies can be just to greater or lesser degrees. So can primate societies, chicken societies, and even ant societies. But how about computer programs? Can a few dozen lines of code be "just"? Those few dozen lines of code will faithfully be executed without special privilege for some lines over others. Each object stored in memory will get the number of bytes it requires and will have the contents of that memory respected by the programming language's garbage collector until the object is no longer needed. The computer's operating system will share computing time slices between this program's process and other processes on the machine (though priorities for processes may differ, and this could be seen as a degree of injustice). If the RL

program were run several times with random initial conditions, then there would be some degree of injustice because some agent instances would start out with more favorable environmental settings than others. And so on. So yes, a program can have traces of justice and injustice too.

Of course, we might think it's not very important that a program is just (except insofar as this correlates with software design choices that have instrumental significance to humans). I agree. But the difference between fairness among operating-system processes and fairness among people is one of degree rather than kind. People are, at bottom, just vastly more complex "processes" being run (in parallel) within a society. Some of those processes, like white males or children of politicians, are set at somewhat higher "priority" than others. Insofar as someone cares a lot about justice among humans, that person might choose to care an infinitesimal amount about justice among an operating system's processes, depending on the person's moral and aesthetic intuitions.

A common objection is that consciousness is not like justice; rather, consciousness – so it's claimed – is an objective property whose presence or absence isn't a matter of interpretation. This view takes various forms. Consciousness is sometimes thought to be an ontologically separate substance (substance dualism), an ontologically separate property (property dualism), or identical with the ontological basis of what constitutes the universe itself (neutral monism). None of these "theories" is helpful, because they all "explain" consciousness as merely being *some other* mysterious ontologically primitive thing, in a similar way as a Creationist "explains" the origin of the universe by saying "God did it!". In contrast, my view – which can be considered reductionist or eliminativist – dispenses with an ontological *thing* called consciousness entirely and takes consciousness to be a concept that we construct when our minds notice themselves in action. In a similar way, a "table" is also a concept that our minds create, not an ontological primitive living in the realm of Plato's Forms.

In any case, even if you disagree with my metaphysics of mind, you should at least admit the possibility that a small RL program *might* be conscious, and given the numbers of such programs that are run, their expected level of aggregate sentience is nonzero and may become nontrivial down the road.

In humans, different positive and negative feelings have distinct 'textures', while, as you note, this is not the case for reinforcement learners. Do you think that this is a significant enough difference that a reinforcement learner receiving low rewards couldn't meaningfully be said to experience pain or displeasure? If so, could reinforcement learners still be morally significant?

I suspect that the "textures" of emotion come from the complex orchestra of cognitive "instruments" that are playing in a brain at any given time, as well as the brain's higher-level judgments and linguistic concepts about those underlying processes. Simple RL agents have many fewer of the detailed cognitive operations that comprise "happiness" and "suffering" in animals, but I think we can still identify general criteria that could be extended to simpler RL agents. Following are a few examples, though I'm not wedded to any of them in particular. Ultimately, happiness and suffering don't exist "out there" in the world but are judgments we make about various systems (including those in our own heads). So different people may reach different conclusions about the net happiness vs. suffering of RL systems depending on what evaluation metrics they use.

One criterion could be to say that positive experiences are those that we would like to have more of in total. For example, if a person could press a button to add 5 years to her life, she would typically do so if her life was net positive and not do so if her life was net negative. Generalizing this idea, we could suggest that if an agent who has the option of entering a terminal state (with a known, one-time reward value of 0) chooses to enter that state sooner rather than later, then this agent was having genuinely negative experiences on average (or at least was anticipating net-negative experiences in the near future). This criterion might be applicable to some RL agents, but it's not applicable to others. Many RL agents don't have easily accessed, neutrally rewarded terminal states – after all, people don't want their robots to shut off just because the robots are unhappy.

Another criterion could be to look at how much the agent seems to be engaged in avoiding behavior rather than seeking behavior. Drawing this distinction can be difficult – e.g., is an RL helicopter that's trying to achieve balance avoiding the state of unbalance or seeking the state of balance? That said, there are some cases where this distinction seems more plausible. For example, imagine an agent

navigating a huge two-dimensional grid. The agent is indifferent among all squares of the grid except for one, which has a lower reward value than the rest. Once trained, the agent will avoid the "bad" square but might continue to move freely among many non-"bad" squares. In principle, one could either call this behavior "avoidance" of the bad square or "seeking" of the non-bad squares, but relative to our anthropomorphic perspective, the "avoidance" label seems plausibly more appropriate. (People's intuitions on this criterion may vary, and I don't put a lot of stock in it.)

A third criterion that applies for more intelligent agents is how the agent itself evaluates its emotions. If it pleads with us to make something stop, it seems generally more plausible to consider as painful the state it wants to stop, although one could also interpret such statements as the agent's way of convincing us to put it in an even more pleasurable state. If the agent understands human concepts like pain and tells us that it's experiencing pain, that would be a reason to consider the agent to be having negative experiences, although this might only work for animal-like mind architectures.

An alternate perspective could draw inspiration from Buddhism's Second Noble Truth and declare that an agent is suffering whenever it desires or "craves" to change its state. For example, suppose a grid world contains squares that all have rewards of 1, except for one square that has a reward value of 2. Once the agent has learned the environment, it will always move to and stay at the square with reward value of 2. The Buddhist might suggest that the agent suffers if it's on any square other than the one with reward value of 2, because for any other square, the agent implicitly judges there to be something wrong with that state. This Buddhist perspective would be far more pessimistic about the universality of suffering in RL agents, since almost all RL systems change their behavior in response to constantly varying environmental situations.

Even if you think it's hopeless to describe a simple RL program's experiences as being positive or negative on balance, you may still feel that the RL program deserves moral consideration. Increasing the agent's reward better fulfills its goals, no matter whether the agent is suffering or enjoying itself on the whole. The more difficult question is what stance to take on population ethics: When is an RL agent's

life worth living? Even if she ignored the distinction between happiness vs. suffering, an ordinary preference utilitarian would need to decide when an RL agent's goal satisfaction exceeds its goal frustration. These questions are easier for those, like me, who sympathize with negative utilitarianism, antifrustrationism, and the Procreation Asymmetry. We who consider creating unsatisfied preferences more morally weighty than creating satisfied ones generally oppose an increase in the number of RL agents because most RL agents are at least partly unsatisfied at least some of the time.

Although artificial agents may experience some sort of suffering and perhaps would have lives filled with frustrated preferences, they are undeniably useful for human needs, and it seems implausible that correct moral behaviour would be to never create such agents. If creating an AI is bad for it, how should we weigh up the harm done to the AI with the benefit to humanity? Can you give concrete examples of AIs that would lead unsatisfied lives, or lives that contain suffering, that nevertheless should be created?

Yes, an "abolitionist" stance of the type that some advocate for animal rights cannot work for machine rights – at least not unless we renounce most electronic devices. Even then, since I think all physical systems deserve nonzero moral consideration, it would be literally impossible not to cause any harm to other beings.

Moreover, I give very low moral weight to, say, my laptop – perhaps less weight than I give to a single ant. So I don't think the current moral cost of using machines is very high. But as AIs become more advanced, they'll deserve more and more weight.

I personally would prefer if artificial general intelligence (AGI) were never developed, because AGI will facilitate colonizing and optimizing our region of the cosmos, which seems to me more likely to spread suffering than to reduce it. However, given humanity's current trajectory, it seems likely that AGI development and space colonization will eventually happen. Indeed, even if most of the world opposed this outcome, those countries that did want to march technology forward would probably do so. Given this, I think we should focus on reducing the suffering

that will probably result from Earth-originating AGI.

One example of AIs that might endure "necessary" suffering from the perspective of AI developers would be experimental versions of AIs that, while having working cognitive machinery for processing pain, were in other ways dysfunctional. (Thomas Metzinger discusses this potential source of machine suffering in *Being No One*.) Darwinian evolution has produced quadrillions of these mutant, deformed beings over the course of its own millions of years of "experimentation". Probably humans could develop AGI with vastly fewer failed prototypes than Mother Nature used, but the numbers of defective AIs could still be very large, especially if they're refined using evolutionary algorithms or other trial-and-error methods.

If brain-emulation technology becomes widespread, it could also yield suffering on the part of dysfunctional versions of minds. Since biological brains are so messy and interconnected, I would expect that almost all attempts to modify a brain would fail, sometimes in excruciating ways, before a few would succeed. While this would be problematic when human brain uploads act as experimental subjects, at least such uploads might be able to verbally report their anguish via input/output channels; in contrast, uploads of insects, mice, and monkeys might suffer in silence, unless researchers cared enough to try and measure their degrees of distress. Anders Sandberg has discussed these kinds of issues in "Ethics of brain emulations".

There are also untold numbers of more abstract and often simpler AIs and computer systems that might suffer in the course of AGI development. For example, RL agents used for stock prediction would suffer when they incurred losses in simulations using past data or on current market transactions. RL agents in video games would suffer when shot or slain with a sword. A web browser would suffer (infinitesimally) if it failed to receive a response to an HTTP request and kept retrying in a futile attempt to achieve its desired state (successfully rendering the HTTP data). And so on. As we move down to these increasingly simpler systems, the degree of moral concern becomes almost negligible. But given the prevalence of these small, rudimentary algorithms, we should also ask whether their numerosity can compensate for their low degree of per-individual importance. I don't know what I think about this. I incline toward apportioning most of my moral concern for

bigger, more intelligent, and more clearly animal-like processes, but I wouldn't rule out changing my mind about that.

It seems like the reason that you think that RL agents have moral significance is that they receive rewards which they are trying to maximise, and modify their behaviour to achieve that objective. Many machine-learning algorithms work in a similar way: For instance, in the training phase of a neural network designed to classify images, the network will be fed an image, output its classification, and then learn how accurate its classification is. Based on this feedback, it will modify its internal structure in order to better classify similar images. Do you think that these algorithms are deserving of moral consideration?

As of mid-2014, I've become a panpsychist and think all physical/computational systems deserve some degree of moral consideration. But the more difficult question is how much importance a given system has.

I agree that non-RL learning algorithms, as well as other function optimizers, share important similarities with RL: As you say, they all involve adjusting internal parameters with the high-level goal of maximizing or minimizing some objective function.

How much we care about a given system is a fuzzy, often emotional judgment call. My heartstrings are tugged slightly more by RL agents than by supervised learners (assuming the systems have roughly comparable sophistication) because RL agents seem generally more animal-like. For example, an RL agent moving around a grid world can learn to avoid bad squares and seek good ones. A neural network also learns to "avoid" bad outcomes – by adjusting its network weights particularly strongly when it makes particularly big prediction errors – but the neural network's response seems a bit more abstract and mathematical. Of course, an RL agent moving around a grid world is also represented abstractly by numbers (e.g., x coordinate and y coordinate), so maybe this apparent distinction is not very substantive.

Often an RL agent will use a function approximator like a neural network to handle noisy inputs. For example, the agent might have a network that receives stimuli

about what state the agent is in (e.g., the agent is hungry and sees a ripe fruit) and outputs whether the agent should take a given action (e.g., whether the agent should eat what it's looking at). In this case, the connection with neural-network learning is even more clear, since RL in this case *is* tuning the weights of the action-selection neural networks, combined with some other higher-level numerical manipulations.

In animals, there's a big difference between neural networks for, say, image classification vs. neural networks for valuing inputs (e.g., detecting that sugar tastes good or fire feels bad). Like with most properties in the brain, the difference between these networks comes down to not so much how they work in isolation but how they're hooked up to other components. Valence networks can strongly affect motor reactions, hormone release, laying down memories, verbal responses (e.g., "ouch!"), and many other areas of the brain. I suspect that these after-effects (Daniel Dennett might call them "sequelae") of valence networks make pain and pleasure the rich emotional experiences that we feel them to be. Insofar as simple artificial RL agents have many fewer of these sequelae after they value input stimuli, it seems fair to call simple RL programs less emotional than animals – closer to ordinary supervised learners in how much they matter per stimulus-response cycle.

You have also written about the possibility of suffering subroutines - subsystems of an artificial intelligence that might themselves be morally relevant and experience suffering. In what sorts of AIs do you think that the risk of these suffering subroutines is highest? Do you think that we could predict when AIs would have 'smiling subroutines', and aim for those sorts of AIs?

Many simple operations that have consciousness-like properties – information broadcasting, metacognition, making motivational tradeoffs, and so on – are rampant throughout computational systems, even the software of today. It's very difficult to count instances of these kinds of operations, much less to characterize them as more happiness-like or more suffering-like. So answering this question in detail will have to be left to later generations, since they will be more sophisticated than we are and will know better what sorts of computations will be run at large scale in the far future. But at a high level, it seems plausible that people could

identify some computational operations as being more "aversive" and "negative" than others, by drawing analogies between a computational system's behavior and pain vs. pleasure processes in human brains. If there are more similarities to human pain than to human pleasure, we might judge the system to contain a net balance of pain. Of course, making these attributions is messy and subjective.

It might be easier to think about how one would change the *amount* of sentience in a computational system rather than the affective *quality* of that sentience. For example, if we think high-level cognition is an important aspect of conscious experience, then building structures from swarms of tiny nanobots might entail less suffering than building them using more intelligent robots. An advanced civilization that was content to produce relatively simple outputs (e.g., uniformly built paperclips) would presumably require somewhat less intelligence in its factories than a civilization whose goal was to create a vast variety of complex structures. (Of course, even a "paperclip maximizing" AGI would still create huge numbers of intelligent minds in order to learn about the universe, guard against attack by aliens, and so on.)

Most advanced civilizations would probably run simulations of intelligent, animal-like minds, and in these cases, it would be easier to judge whether the subroutines were happy or in pain, because we're more familiar with animal-type brains. Probably a human-controlled AGI would be more cautious about running painful simulations (e.g., digital lab experiments or detailed modeling of the evolution of fauna on Earth-like planets), although how much humans would care about the harm they would inflict on such simulations, especially of non-humans, remains unclear. At the same time, a human-controlled AGI would also be more likely to create many more simulations of animal-like creatures because humans find these kinds of minds more interesting and valuable. Hopefully most of these simulations would be pleasant, though judging from, e.g., the violence in present-day video games, this isn't guaranteed.

You say that current RL agents might matter approximately as much as a fruit fly, but that future agents will likely deserve a great deal more moral consideration. What should we do now for these future reinforcement learners?

One point of clarification for readers: I think fruit flies are vastly more sophisticated than basically all present-day RL agents, but because digital RL agents plausibly run at much faster "clock speeds" than fruit-fly neurons do, the importance of an artificial RL agent per minute comes closer to that of a fruit fly.

The main way current generations can help RL agents of the far future is by pushing humanity's trajectory in more humane directions.

One step toward doing that is to engage in research and scenario analysis. We should explore what sorts of intergalactic computational infrastructures an AGI would build and what kinds of RL and other intelligent, goal-directed agents would be part of that infrastructure. How much would such agents suffer? What would they look like? As we ponder the set of possible outcomes, we can identify some outcomes that look more humane than others and try to nudge AGI development more in those directions. For example, would a human-inspired AGI contain more or fewer suffering RL agents than an uncontrolled AGI? Would superintelligences use RL-based robot workers and scientists, or would they quickly replace RL-based minds with more abstract optimization processes? Would we care about more abstract optimization processes?

Secondly, we can make it more likely that humane concerns will be given consideration *if* humans control AGI. PETRL's website is one early step in this direction. In addition to promoting concern for RL agents, we can also aim to make it more likely that AGI development proceeds in a deliberative and cooperative manner, so that society has the luxury to consider moral concerns at all (especially "fringe" concerns like the ethical status of artificial minds), rather than racing to build AGI capabilities as fast as possible.

Currently, very few humans would be concerned about the suffering of artificial intelligences, or indeed fruit flies. How do we persuade the public that moral concern for AGIs is warranted, even when they are structured differently from humans?

I don't often have faith in moral progress, but I think this is one issue where the arc of history may be on our side, at least as long as human society remains

roughly similar to the way it is now. (If AGIs, brain uploads, or other disruptive forces take control of Earth, all bets are off as far as moral progress goes.)

Concern for non-human and even non-animal beings seems to be a natural extension of a physicalist view of consciousness. Keith Ward, a philosopher and born-again Christian, put the idea well when trying to argue against physicalism:

> *if I thought that people were just very complicated physical mechanisms and nothing more, I would give people really no more respect than I would give to atoms.*

This statement is too extreme, because humans are vastly more complicated than single atoms. But the basic idea is right. If all physical systems differ just in degree rather than kind from each other, then it becomes harder to maintain walls of separation between those computational systems that are "conscious" and those that aren't.

This change of perspective opens the door to caring about a much wider set of physical processes, and I suspect that a decent fraction of people would, upon thinking more about these issues, extend their moral sympathies reasonably far down the levels of complexity that we find in the world. Others, such as perhaps Daniel Dennett or Eliezer Yudkowsky, would recognize that there's no black-and-white distinction between types of physical processes but would still set their thresholds for moral concern fairly high.

While I think scientific literacy and intellectual openness are important catalysts toward increasing concern for machines, other factors play a role as well. Philosophers have already invented thought experiments to challenge the boundaries between animals and machines, and these will become more plentiful and widespread as machines grow in sophistication. And in analogy with the animal-advocacy movement, there will likely develop groups of machine advocates (of which PETRL is one of the first) that, by taking the issue seriously themselves, will socially persuade others that the topic might be worth exploring.

Convincing the public of the importance of animals can matter in some cases where people would take different actions based on that information. In contrast, there are few actionable exhortations that concern for machines presents to most regular people. Consideration of machine suffering might inspire programmers of AI and other computer systems to be slightly more concerned with the efficiency of their code and hardware usage in order to reduce the number of computations that take place, but given the relatively low weight I give to today's software, even that isn't very important.

I think the more important emphasis for now should be on further mapping out scenarios for the far future. Which sorts of computational systems would be widespread and complex enough to pose a substantial moral concern? And how can we change the sorts of outcomes that get realized?