The Value Function of Human-Compatible AI

Jongmin J. Baek
UC Berkeley

May 15, 2016

1 Motivation

It seems that complexity theory has a surprising relationship with morality. Generally, if a decision renders the environment seemingly more *complex* than another decision, this decision is considered moral. We seem to respect or otherwise see value in complex things, such as humans, works of art, and R2-D2. In his book *I Am a Strange Loop*, the cognitive scientist Hofstadter quips,

> In one of the *Star Wars* films, I recall seeing a huge squadron of hundreds of uniformly marching robots - and when I say "uniformly", I mean *really* uniformly, with all of them strutting in perfect synchrony, and all of them featuring identical, impassive, vacuous, mechanical facial expressions. I suspect that thanks to this unmistakable image of absolute interchangeability, virtually no viewer feels the slightest twinge of sadness when a bomb falls on the charging platoon and all of its members - these factory-made "creatures" - are instantly blown to smithereens. After all, in diametric opposition to C-3PO and R2-D2, *these* robots are not creatures at all - they are just hunks of metal!... What is it, then, that gives us the undeniable sense that C-3PO and R2-D2 have a "light on" inside?

I wish to run with this quote in the following direction: the uniform robots seem evil because they have little to no complexity, i.e. can be described in a few simple lines of a formal language, whereas C-3PO and R2-D2 seem "soulful" because they are more complex and unpredictable. From here, building an ethical, human-compatible AI reduces to building an AI with a value function such that it, at the least, maintains and, at the best, maximizes the complexity of its environment. I wish to present a sketch of such a Human-Compatible Agent and apply it to a toy scenario with a formidable moral dillema.

2 Clarifications

Given the claim that a moral decision is a decision that renders the environment more complex, I have encountered a number of objections. A most common objection is that the complexity of any environment, when considered down to the molecular level, never changes; for example, the complexity of a dead cat is equal to the complexity of an alive cat, for the same physical laws apply to the same number of molecules. But when I talk about the complexity of an environment, I really mean the complexity of the *representation* of the environment; there is a certain abstraction barrier in this representation, as humans are wont to have. What I mean is that no human being conceptualizes his environment as a set of molecules. Instead, he uses the five senses. We can give our Human-Compatible Agent a similar state representation; for example, the representation of a cat, dead or alive, can be a video of the cat shot with an iPhone 6S.

3 Russell's Cat

Russell proposes the following problem: if the parents are out and the children are hungry, what would stop a house-keeping artificially intelligent agent from cooking the cat and serving it to the children? My answer is simple: an alive cat is more complex than a dead cat, ergo environments in which a cat is alive is more desirable than environments in which the cat is dead.

Let us formalize the problem of Russell's Cat. We define the house-keeping agent with a set of states $s \in S$, a set of actions $a \in A$, a probabilistic transition function $P : S \times A \to \mathbb{R}$, a value function $V : S \to \mathbb{R}$, and a reward function $R : S \to \mathbb{R}$. In each step, our agent is in some state, which constrains his set of available actions. Our agent then chooses the action that maximizes his value function. The value function is a function of the current observation and an action taken. It is the sum of the values of all possible futures, weighted by their respective probabilities, and weighted also by a discounting factor γ:

$$A(s) = argmax_{a_i} \sum_{s_i} P(s_i|s, a_i)V(s_i)$$

$$V(s) = R(s) + \sum_{s_i} P(s_i|s, A(s))(\gamma V(s_i))$$

$$R(s) = K(s)$$

Where K is the Komolgorov Complexity of some state. Now consider the fact that a state with an alive cat has more Komolgorov Complexity than the same state with a dead cat; it seems intuitively obvious that it takes fewer lines of code to describe a state devoid of a furry and adorable feline as opposed to a state containing one. From this we can easily deduce that our housekeeping agent will definitely not kill the cat, instead opting for the safe, already dead pieces of meat in the fridge. Cooking dead pieces of meat does not decrease the complexity of the state as much as killing an alive cat does.

This formulation has the advantage of being more robust than an arbitrary rule such as "whatever happens, don't kill the cat to feed the children." What if the parents never return from their night out, days pass, the fridge is empty, and the children are starving? Our agent should be able to compute the value of a future in which 1) the children are dead and the cat is alive, and in which 2) the cat is dead and the children are alive. Assuming that the children are more complex, or have the potential to be more complex, the Human-Compatible Agent will kill the cat to serve to the children in such dire circumstances. Of course, if calling the police is an available action, the Human-Compatible Agent wil predict that a future in which the police save the children (and the cat) is more complex than a future in which the cat is dead, and as such, choose that action.

4 Good Children, Bad Children

There seems to be a paradox here. If the value function of morality is truly to increase the complexity of the state, is it immoral to read a book instead of smattering it against the window? Is it immoral to have world peace as opposed to world war? I assert that this seeming paradox is an effect of the flexibility of our discounting factor γ. In a well-known study, children were presented with a brownie and a promise of three brownies if the brownie were not consumed for five minutes. The children who managed to wait were much more successful later in life than the ones who failed this agonizing test. Let's assume the successful children were more moral, pour them into the mold of our model, and examine their discounting factor γ.

Consumption of brownies, of course, cause an upsurge of complexity in the (internal) state, and is therefore desirable. Let's assume linear increase in complexity of internal state as the number of brownies increase. Let's further assume that there are just three states: s0) no-brownies-consumed, s1) one-brownie-consumed, s2) three-brownies-consumed. Let's assume there are two actions possible, a0) eat or a1) not eat. The world is deterministic. All children start at s0. The failed child chooses s1 over $\gamma \cdot s2$.

Failed child:

$$
\begin{aligned}
a(s0) &= argmax_{a_i}(P(s1|s0,a0)V(s1), P(s0|s0,a1)V(s0)) = a0 \\
V(s1) &> V(s0) \\
V(s0) &= R(s0) + \gamma \cdot (V(s2) = R(s2)) \\
V(s1) &= R(s1) \\
R(s1) &> \gamma \cdot R(s2) \\
x &> \gamma \cdot 3x
\end{aligned}
$$

Successful child:

$$
\begin{aligned}
a(s0) &= argmax_{a_i}(P(s1|s0,a0)V(s1), P(s0|s0,a1)V(s0)) = a1 \\
V(s0) &> V(s1) \\
V(s0) &= R(s0) + \gamma \cdot (V(s2) = R(s2)) \\
V(s1) &= R(s1) \\
R(s1) &< \gamma \cdot R(s2) \\
x &< \gamma \cdot 3x
\end{aligned}
$$

In this model, the patient factor, γ, plays a big role in making moral decisions. We seem to believe that having more patience, or a numerically bigger γ, makes us more moral. Intuitively, γ is probabaly an approximation of the amount

of dread the agent feels: if the agent feels that he is going to die soon, he will want to maximize utility as much as possible as soon as possible, gleaning cheap pleasures before he dies. This will lead to a heavily weighted γ. On the other hand, an enlightened agent, unafraid of death, will consider the bigger horizon, choosing decisions that may have little effect in the short term but will maximize complexity in the long term. This will lead to a lightly weighted γ. It seems that we would want to set γ to be close to 1 for a Human-Compatible Agent.

5 Conclusion

I have described a brief sketch on how complexity theory may be used to approximate morality. Specifically, I asserted that a value function that is set as the complexity of the environment with lightly discounted future rewards leads to morally behaving agents. I wish to further investigate the parts where I simply waved my hand instead of providing a rigorous proof. A big problem is that the complexity of the environment is probably uncomputable. Perhaps some approximating algorithms can be used instead. Moreover, it is ambiguous how one would go on about using such a generic value function for specialized tasks. Perhaps the value function of increasing the complexity of the state can be used as an add-on to a preexisting value function, pulling the breaks only when the agent is about to do something severely immoral, i.e. severely decreases the complexity of the state, such as killing a cat.