Maximizing Possible Futures: A Model For Ethics and Intelligence

This paper argues for the use of a specific utility function: maximizing possible futures. The concept of ‘maximizing possible futures’ was originally proposed by Physicist Alex Wissner Gross as a model for intelligence. We agree that Maximizing Possible Futures is indeed a good general description of intelligent behavior. In this paper, We argue that it is also a good description of *ethical* behavior.

First, we examine problems with utilitarianism. Next, we present arguments in favor of Maximizing Possible Futures as an ethical system. Finally, we use the concept of “maximizing possible futures” to trace through common arguments for and against utilitarianism. The concept of maximizing possible futures shows both why the intuition behind utilitarianism is sound, as well as why the well understood flaws behind utilitarianism *are* indeed flaws.

Problems with Utilitarianism

What is Utilitarianism?

Utilitarianism is defined as ‘the greatest good for the greatest number’.   In simple terms, this means that ethical actions consist of actions which maximize the total sum of happiness in the world, minimizing the total sum of sadness. How to weigh these two against each other (increasing good vs. decreasing bad) is not agreed upon.

Hedonistic Utilitarianism defines good in terms of pleasure and bad in terms of pain. Other forms of utilitarianism allow for arbitrary ‘utility functions’ which can differ from agent to agent, representing what that specific agent values.

Allowing agents to define their own utility functions has some benefits. Hedonistic Utilitarianism can lead to the ethical equivalent of junk food: things that feel good, but which we know are not good for us. For example, if everyone on earth consumed a large amount of heroin, we’d all feel great, and then die, at which point we’d feel no pain. Intuitively, this would not be a good outcome, but Hedonistic Utilitarianism says it would be.

Utility Functions are Arbitrary

If we allow agents to have arbitrary utility functions, we can get into other problems. Detaching ethics from emotion does free us from the absurd conclusions detailed above, but it can lead to other problems.

What if most people enjoy fighting? If we equate ethical behavior with maximizing utility functions, and most people enjoy fighting, then anyone trying to reduce conflict on earth is behaving unethically.

What if we live in a society with a strange antipathy towards anyone wearing purple pants? Would wearing purple pants then become unethical behavior?   If we say that utility functions are *not* simple pleasure and pain, we come to an important question. Can a utility function be "wrong"?

If a utility function can't be wrong, then utilitarianism says that ethical behavior consists in going along with the crowd, whatever the crowd is doing.   This hardly seems like a satisfying answer.

If the answer is yes, utility functions *can* be wrong, then we are now stuck with the problem of determining which utility functions represent ethical behavior, and which ones do not. This is, essentially, the problem of ethics. Utilitarianism hasn't given us a real answer if we agree that some utility functions can be unethical. Is utilitarianism still useful in this case? I argue that it is.

Utilitarianism as Error Correction for Ethics

If each person in the group has their own utility function, and these functions can be wrong, it is reasonable to assume *my* utility function could be wrong. From an epistemic point of view, we might argue that my ability to predict my emotional response to a situation is flawed and imperfect. We know that our ability to predict the future outcomes of actions is flawed and imperfect.

By following Utilitarianism - that is, considering other agents' utility functions in addition to my own - i may be able to *better* maximize my own utility function than if I merely pursue increasing my own utility directly.

Suppose there is some 'correct' utility function. To use the language of moral foundations theory (Haidt), suppose that the 'correct' utility function is expressible as a vector of weighted values in each of the categories. Let each agent's personal utility function consists of some linear combination of a) the correct function and b) a bit of random noise. No one agent's utility function is likely to be correct. How would we know if one of them were? Their stated utility function would not align with ours, and we would trust our own intuition over someone else's, unless we had massive amounts of evidence.

By combining the utility functions of large number of agents, might might be able to cause the noise to cancel out[1]. In other words, perhaps some people value purity *too much*, whereas others value it *too little*, and by combining the utility functions of a large population, we come to the exact correct amount to value purity - and the same goes for the other dimensions in Moral Foundations Theory.

Whose Utility Functions Do We Consider?

It might be tempting to answer "we consider everyone's utility function!" but this is a seriously problematic claim. Although this claim appeals to the value of fairness and egality, it is impossible to take seriously from an epistemic standpoint.   There are 7 billion persons on earth; it would be absurd to claim that i take everyone's feelings into account whenever i act.

A more realistic standpoint might be "My overall utility function considers the utility functions of myself and a few persons in my immediate family. I place a lighter weight on the utility functions of more distant family, friends, and acquaintances. Then there are general terms to account for the utility of faraway persons, of whom i have no personal knowledge,but still care for.

This later formulation is both far more realistic, but intuitively unsatisfying. It doesn't seem like a correct ethics at all, but rather an acceptance of limited epistemic capabilities.

If we start poking around the question of "whose utility functions do we consider", we start to find ourselves asking questions like "is it ethical to kill one human if it says ten thousands cows?" We might reach a point where we consider that the ethical course of action is to kill the human race so that we can reduce the suffering of animals. Even limiting our discussion to humans, we come to the question of people who are mentally ill, and children. If the world has a large population of children, should 'staying up later than bedtime' be considered more ethical, because that is what children would prefer? If we are willing to discount the utility functions of some humans, when does that process stop? Should we weigh the president's utility function equally with everyone else?

We need a way out of utilitarianism, which hasn't really solved the question of ethics. Utilitarianism *is* useful as an error correction method, but it doesn't' solve the fundamental question of what ethical behavior *is*.

[1] There is a relationship here to the general pattern in quantum computing: a problem is encoded in a quantum system so that all failed solutions cancel each other out, leaving only a correct solution remaining, if one such solution exists.   So if each person has a randomly assigned vector of preferred moral values, and we find ourselves living in a messy dirty world, but there is this *one person* who seems to be consistently in a manner we all agree is ethical, we would have discovered true ethics via quantum computing.

This paper proposes a utility function that could be a ‘correct’ ethical system. That utility function is ‘maximizing the possible futures’.

What is “Maximizing Possible Futures?”

“Maximizing possible futures” has been proposed as a general model of intelligence. An agent that maximizes possible futures is an agent that models its environment, computes the possible states accessible to that environment in the future under various courses of action, and selects a course of action which leads to the most possible states of that environment in the future.

For a simple example, consider a computer program that plays chess by making the move that gives it the most freedom to move in the future. If one move leads to a board configuration that gives the white player 10 possible moves, and another move leads to a board configuration with 30 possible moves, the agent would choose the second move. This is a well-known approach for AI systems that play chess.

This theory was originally proposed as a theory of *intelligent* behavior. This paper argues that maximizing possible futures is also a framework for *ethical* behavior.

Intuition

“Maximizing possible futures” contains a lot of things that we would consider intuitively ethical.

- Encouraging Learning and Pursuit of the Truth

  An agent that works to maximize possible future states accessible to its environment must have the ability to accurately model its environment. It must have the ability to accurately assess the impact of its choices on the future states accessible to its environment.

- Encouraging Environmental and Self preservation.

  If the agent itself were to be destroyed, any future states involving that agent in any non-destroyed state would become inaccessible to the system. This leads to a natural sense of self preservation: Any agent that operates under an ethical framework of maximizing possible futures would take steps to keep itself alive - unless these steps would lead to a reduction in future possibilities for the environment in which the system operates.

Maximizing the future states available to a system leads to an agent that tries to protect the system it inhabits. An agent operating on this principle might take action that would lead the agent to be destroyed, if it believed that this action would lead, ultimately, to the preservation of the system the agent inhabits.

- Encouraging Health and Growth

  A healthy person can have more experiences and do more things than a sick person. An educated person can have more experiences than an uneducated person. A strong person can engage in more activities than a weak person.

  Any agent whose ethical framework is “maximizing possible futures” of the system it inhabits will act in a way that encourages the health, growth, and development of all other agents in the system it’s part of.

  Of course, if this agent observes a virus that is destroying lots of healthy species, the agent would not encourage the growth of the virus, because the total number of states accessible to the system decreases as the virus grows.

- Discouraging Death and Destruction

  When a building is destroyed, all states of the environment where that building has children playing in it are no longer accessible.

  When a person dies, all states of the environment where that person is alive and well are no longer accessible. An agent attempting to maximize the possible future states accessible to the system it inhabits would attempt to limit death and destruction wherever possible.

  If an old aspect of the system has lots of resources which could lead to more possible configurations elsewhere, then an agent following the ‘maximize possible futures’ ethical model might allow the old aspect to die, or even take action to destroy it: the agent would evaluate the future states accessible to the world, and move towards a world of more possibility.

- Encouraging Exploration

  A species that lives on one planet has far fewer accessible states than a species which lives on hundreds of planets. An agent that operates to maximize possible states accessible to the system it inhabits would work to encourage exploration and settling of distant environments.

- Ego-Loss

  An agent which operates to maximize the possible future states of the system it inhabits only values itself to the extent that it sees itself as being able to exhibit possible changes to the system, in order to maximize the future states accessible to it.

  In other words, an agent that operates to maximize possible future states of the system is an agent that operates without an ego. When this agent encounters another agent with the same ethical system, they are very likely to agree on the best course of outcome. When they disagree, it will be due to differing models as to the likely outcomes of choices - not on something like core values.

Argument for Maximizing Possible Futures

Utility Functions as Models

What is a utility function? Where do they come from? If we reject the hedonistic model and say that utility functions can be arbitrarily stated by the agents, we can say no more than tautologies.

If we consider hedonism as a base case, then any meaningful answer must ultimately come to evolution. We derive pleasure from food because food helps us survive. We derive pleasure from sex because it helps our species survive. If we didn't enjoy eating, and starving weren't painful, we might not bother to feed ourselves. If raising children didn't feel so satisfying, and having sex weren't so pleasurable - we might not bother to reproduce. Another species would have taken our place.

I argue that our utility functions should be seen as *models.* They should be seen as the result of millions of years of genetic evolution selecting for genes that *maximized their chance of existing in the future.*

In other words, our genes have been attempting to maximize the possible future states accessible to them. This is not an argument that genes are conscious, or that they could be described as having intentions. This is an argument that the choices which make human beings feel good tend to line up, *roughly,* in the same direction as the intuitive values of "maximizing possible futures". It's an argument that *hedonistic utility functions are models.*

We crave sugar for evolutionary reasons. It used to be a lot harder for us to obtain. Obtaining sugar gives us energy, which gives us more capacity to act on our environments.

If we see our utility functions as being *models* of (local views of) the correct ethical system*,* we see validity of the intuition behind utilitarianism. When you have lots of different biological systems modeling “the approach which increases future possibility”, moving in a direction that all of those models say is correct isn’t a terrible idea.

Yet if you *forget that you are dealing with models*, you might move in a direction that plays to the bias in the models, rather than in a direction you actually want to go in. You might get the ethical equivalent of junk food: actions that increase utility while harming the world.

Suppose we humans are trying to build a tower to the heavens. The tower has gotten so tall, we can’t build a device to measure the tower. We don’t know whether we are making progress, so we’ve placed men and women at the base of the power, to try to estimate the height of the tower, using the motion of the stars or other tools at their disposal. If we become too focused on trying to change the estimates, we might feel like we are making progress towards some good, while really we are further driving our models of reality of whack.

Utility functions are *feelings,* and while feelings are useful inputs, they are not reality.

Conclusion

Much of modern ethics is expressed in terms of utilitarianism. Although it has known flaws, it is widely considered to be the best model there is. The I hope this paper provides an argument that “Maximizing Possible Futures” is an excellent such utility function.