Artificial intelligence and utility : deep problems for A.I. ethics

Trevor Howard Groves
University at Albany, State University of New York, tgroves1221@gmail.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Ethics and Political Philosophy Commons

Recommended Citation
https://scholarsarchive.library.albany.edu/legacy-etd/2293

This Master's Thesis is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive.
Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Artificial Intelligence and Utility:
Deep Problems for A.I. Ethics

By:

Trevor H. Groves

A Thesis
Submitted to the University at Albany, State University of New York
In Partial Fulfillment of
the Requirements for the Degree of
Master of Arts

College of Arts and Sciences
December 2019
Abstract:

As technology advances further, new problems arise for us to contend with philosophically. Indeed, popular media has already begun presenting the possible issues our society must deal with. One of these is A.I., and how they will challenge old notions of subjects like identity, consciousness, and ethics. It is this third category that this paper is concerned with, namely how the existence of A.I. will affect our choice of ethical theory in regards to their treatment. It is my intention to demonstrate that one particular category of ethical theory, hedonistic consequentialist ones, will be ill equipped to handle these considerations. More specifically, A.I. are likely to give us new considerations of old problems that will be very difficult to overcome.
Section 1: A Brief Overview of an Old Problem:

For consequentialist ethical theories, particularly hedonistically focused ones, there has been posed the hypothetical of a “Utility Monster”. These are beings who have deep, ravenous, and unending appetites for any good in consideration for distribution, such that they experience greater amounts of utility than what anyone else would receive. Such a being would naturally bias any and all considerations towards their inclinations, regardless of the deficit caused in others. Given that we often care about who gets something and how many, as much as we care about why, this can obviously lead to unwanted results.

For example, suppose we have the decision of giving a cookie to either Jones or Smith. Jones has an incredibly strong sweet tooth, whereas Smith has a milder, but still present desire for the cookie. Under a Utilitarian account, since Jones has the stronger preference, we give it to them. Then we present another cookie and evaluate who would want it more. Again, Jones wants it more, and this repeats on and on until we have no more cookies left. Jones is stuffed to the gills, whereas Smith is left without even so much as a crumb.

This scenario tends to strike us as somewhat unfair. Even if Smith doesn’t get as much pleasure from a cookie as Jones, it does not seem right to give every cookie to Jones without at least something for Smith, even if Jones would like it more. Normally, utilitarians have an easy out to this through relatively mundane observations. Humans, and in fact, all beings with an appetite in general, have desires that are responsive to their satisfaction. Giving Jones a cookie diminishes her desire for another one. At some point, practically, there becomes a tipping point where Jones’ desire for one more cookie is less than Smith’s. This gives us a “failsafe” of sorts to ensure that there will be at least be a reasonable distribution of resources over time and mitigates the problem.
This is a reasonable defense that prevents the spiral of consumption from being unending and concurs with our everyday observation. We all get sick of something from repeatedly indulging in it, be it listening to a song, eating tasty food, or even participating in a particular hobby. But the insidiousness of the utility monster, in part, is that its desires are completely, or nearly completely unresponsive to its satisfaction in regards to diminishment. Imagine being as thirsty as you possibly could, and while you drank a glass of water and felt all the satisfaction of rehydration, you still craved more water as if you never drank the first glass or any others after it.

The utility monster takes this desire, and pushes it even further, by making its happiness so great from receiving such things like water, food, or even entertainment, that it could outweigh even the desire of dehydrated or starved persons. The utility monster becomes a glutton that demands all our resources, leaving none for ourselves by the sheer magnitude of its satisfaction compared to our own when it acquires such things. Utilitarianism, by the virtue of trying to maximize utility, would be compelled to state that such a case would be the greatest distribution of resources possible, even as the world languishes in destitution.

It might be morally required that we give everything to the person or thing experiencing such great utility from receiving them. Nozick famously described the concept of a utility monster when he presented the original argument in Anarchy, State, and Utopia. Nozick writes “Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater sums of utility from any sacrifice of others than these others lose" (Nozick 1974, p. 41). In which case, "the [utilitarian] theory seems to require that we all be sacrificed in the monster's maw, in order to increase total utility” (Nozick 1974, p. 41).

In other terms, a utility monster gains such great utility from any particular thing that it outweighs the loss of someone having to sacrifice it to the monster. Thus, we become obligated
to forego everything for ourselves, in order to give it to the utility monster so that we may bring about the greatest utility.

The response to this objection is essentially the one we gave above in regards to Jones and Smith. It appears that the utility monster poses no significant threat to the theory in practice. It seems difficult to imagine a realistic scenario where the strength of the desire is completely or almost completely unresponsive to whether the object of that desire has been provided. A desire for water diminishes as one receives water to quench thirst, a desire for food decreases as one’s eats and satisfies one’s hunger, etc. In fact, it seems very strange to think of a desire that doesn’t decrease in strength as it is satisfied, and any instance of such a case would very likely be few and far between, if it any case exists in the first place. So in this case, it seems like the concept of a utility monster is relegated to the realm of abstraction and so isn’t a particularly threatening objection, because once we gave some resources to it, it seems that it would eventually stop deriving so much utility from any additional resource.

Even more damning is that it seems unrealistic to think that the desire of an individual could be so strong that it could override every other person who could desire any particular thing. The sheer strength of desire the utility monster would have for any given item would have to be almost inconceivably immense.¹ Not only would the desire have to not diminish in the face of satisfaction of it, but it would have to be so overwhelming that it could override the desires of any existing person and potentially future persons, should we dedicate enough resources to it. What kind of being could reasonably exist that could possess such desires? Even granting that it is in principle possible, it does not appear remotely possible in practical cases to believe such a desire could exist.

¹This particular argument is discussed more at length in (Parfit 1987)
In other words, it appears unreasonable to believe that a utility monster is a credible threat to Utilitarianism, given how difficult it would be to provide a case of one actually existing. A mere threat of conceptual possibility is not grounds for rejection of a theory, unless that threat is likely to materialize in some way. The existence of unicorns would completely rewrite our knowledge of equines, but if none exist, why take them as a reason to abandon our current theories? It becomes the onus of the one proposing the objection to show a practical case of one existing in reality, rather than merely hypothesizing about the mere idea of it being able to exist and provide a challenge to Utilitarianism.

However, notice what these responses claim. Their arguments are not that Utilitarianism has internal mechanisms for dealing with such scenarios even should they exist, but rather, denying their possibility as entities one would have to actually contend with. If an example was provided, these arguments would be unable to resist the disastrous conclusion it would seem to entail from the theory. While other defenses exist for dealing with utility monsters (and we shall deal with those in time), the first step has to be demonstrating that such a case is reasonably possible, and this is where our A.I. enter the picture.

Section 2: Some Conceptual Housekeeping
Before we can begin our exploration of the hurdles A.I. might present before us, we must first understand what kind of thing we are talking about when we say “A.I.” Simple programs that interact with a user’s input and responds accordingly on a computer can be construed as an A.I., likewise, we can also mean something far stronger too. Maybe we’re referring to something more like C-3PO and R2D2 from Star-Wars, beings that intelligently and dynamically interact
with the world, engage in conversations with humans, and in general, seem like they have what can be interpreted colloquially as a mind.

For the concerns with this paper, we shall focus on this latter sense; beings who seem like they are not merely lines of code pre-programmed to respond to various stimuli in pre-packaged ways, but rather, could believably mean to say something like “Please don’t do that to me.” In some way, they’re not just parroting what they were programmed to do, like a chemical reaction under particular conditions, or a rigid conditional of “If A then B”, but what might be characterized in a way like a human’s reaction. There is some individual that is having an actual and aware perception of a stimulus.

Furthermore, in the course of consideration of all these arguments, the A.I. in all these arguments are meant to be construed as more physical ones. We can conceive of A.I. created purely virtually, like a computer simulation. These “Virtual A.I.” create unique and more interesting challenges for us to evaluate. However, I limit myself to more “physical” conceptions of A.I., like the aforementioned droids from Star Wars for a few reasons. The first is that it is easier for us to conceive of such cases. While possible, it is more difficult for us to imagine such beings existing as a program only, without any physical existence. Their radically different composition, one of complete non-corporeality, invites new questions that I do not wish to engage with here.

The second reason is that, in the course of argument, I wish to provide the easier case for consequentialists to deal with, and show how even in the simpler cases, problems abound. Many of the physical limitations a physically realized A.I. would run into would not hold for purely virtual realizations. One can argue from various plausible facts that production of a thousand physical A.I. would run into problems before the production of a thousand virtual ones. Yet if I
can show that these physically realized ones would still present a sufficient problem for consequentialist theories, then the “harder” case of virtual A.I. would only be even more difficult to resist. So for the course of these arguments, I wish to give consequentialists the best possible position to argue from, and show how even that position would be left wanting.

   Next, I shall not go deeply into what it might take for an A.I. to be of a kind that is worthy of being included in a moral theory, or even what would be required for one to possess full-fledged sentience. My approach is one that does not hinge on particular conception or another, but rather, encompasses many possible views on the matter. Of course, our beliefs on moral consideration and sentience must be reasonable. Candidates for reasonable assumptions should resonate with our intuitions, instead of being wildly outlandish and strange. If our assumptions lead us to believe things like rocks, electrons, and puddles on the side of the road are all moral patients, we’ve got bigger problems than just the ones I present.

   To this end, at least some restrictions must be placed upon what kind of theories we can allow into consideration. For example, our theory of A.I. must allow for the possibility of sentience. This restriction has deep ties into our moral intuitions; as has already been mentioned things like rocks and even unaware living organisms like plants and bacteria do not seem like they are worthy of moral consideration outside of their relation to sentient organisms. Their moral worth is determined by how things with some sort of awareness need or use them for their own ends. I commit a moral wrong in killing my neighbor’s garden not because I have wronged the plants contained therein, but rather, because I have upset my neighbor who invested their time and energy into growing it.

   This assumption isn’t a difficult one to believe either, given that, as technology and knowledge in the fields of subjects like information processing progresses, we get better and
better at producing programs that can dynamically interact with the world by performing various
tasks. A.I. can already interact with their user’s voice commands, analyze and process faces in
various conditions, autonomously navigate around objects and manipulate the environment to
accomplish tasks, etc. Of course, none of these are anywhere near sentience, being able to
perform one task very well after being programmed with a suite of tools specifically to enable
them to do that is far and away from being aware. However, given that A.I. development is still
in its infancy, there is still plenty of time and chance for the technology to advance far enough
for us to begin to consider otherwise. Only time and much philosophical debate will tell whether
this assumption will bare itself out, but it is one we must make for the sake of this paper’s focus.

The other assumption we must make is one towards our moral theory. There is a
theoretically infinite amount of possible theories one could adopt, even among just purely
consequentialist theories. Underlying assumptions between these theories come into conflict with
one another and are susceptible to differing problems, so it is necessary to specify what the
parameters are of an acceptable theory. (Shulman et al 2009). Hypothetically, we could come up
with a theory such that it claims that rocks and other non-sentient objects are moral agents. Such
a theory is obviously ridiculous and not worth engaging with, but my arguments would be unable
to engage with them at all, given that they target more usual conceptions of the good, such as
bringing about the greatest amount of happiness, satisfaction of desires, etc for sentient agents.
So for this paper, we shall be considering more “typical” notions that consequentialists lean on
when they say we ought to bring about X.

To put this all together then, for the sake of this paper, we are to assume that sentient A.I
is possible, that they can exist, and that our theories of consequentialism requires that our moral
good require our moral good to in some way to interact with sentience. For if our moral good did
not interact with sentience, not only would it be highly unintuitive, but arguments in general regarding utility monsters would be an incoherent one. I do not believe that most will find this requirement a difficult one to accept, but it beneficial to lay the cards out on the table. Of course, this may somewhat limit the overall scope of my paper, but it is a necessary limitation in order to progress with the problems that I intend to present.

Section 3: A.I. as Utility Monsters

With all this in mind, let us return to the concept of utility monsters. Why does A.I. warrant us to reconsider this problem? I will argue that there are various facts likely to surround A.I. that will serve to make them utility monsters. Granted that these arguments are largely speculative in nature at this point, given that A.I. development is still in its infancy, if even that. I will ground these arguments by proposing relatively modest and believable claims about how A.I. are likely to be like, and why they present these issues as a result. If my beliefs about A.I. do not hold, neither will my arguments, but I am confident that they will reasonably hold to some extent, even if they do not bear out completely.

Starting small with these assumptions then, with the advent of strong A.I. (A.I. with an intelligence comparable to a human’s), there’s an immediate difference about them that places them in an interesting position relative to ourselves, or even organic beings as a class. While humans have been able to manipulate psychological characteristics of animals through continued selective breeding of animals like dogs, there is the possibility of complete psychological control in the creation of an A.I. We completely design its perceptual apparatus, its information processing systems, its very physical parts and components. In other words, we have the deepest
levels of control over both the physiological and the psychological aspects of A.I. at the outset. And this presents dynamics and interactions for consequentialist theories that seem strange and only further contribute to the problem of A.I. as utility monsters.

If our moral obligation is to bring about some psychological good like happiness, and we have complete control (or at the very least, extreme power) over the psychological state of an A.I., then we get the obligation that we ought to make A.I. only experience some form of happiness. After all, how could we be increasing happiness if we made a robot that could be sad, or angry, or anything less than increasing the amount of happiness? And at first, this might not appear to be an entirely bad thing. Why would we make a being and give it the specific ability to suffer? That could be reasonably construed as a cruel act, however, only allowing the possibility of pleasant experiences opens some disturbing scenarios, some of which will be touched upon later in this paper.

For now, this brings us back to the objection that a utility monster is unrealistic. In the case of humans, or even biological organisms as a whole, there is a physical and psychological limit to the capacity one can feel joy. However, it is not reasonably clear that this extends in a similar way to strong A.I. An A.I. does not have to deal with things like the finite supply of dopamine or serotonin in the brain, or the psychological processes that operate within us to dictate how we perceive stimuli. It’s not a given that an A.I. couldn’t live in an endless loop of hedonistic delight, filtered by a perceptual and informational processing apparatus that lets all stimuli be perceived as worthy of causing happiness.

Combined with the previous obligation of ensuring that an A.I. could not suffer and that we ought to maximize utility, we get the problem of the utility monster scenario. A being of potentially near limitless hedonistic experience being able to outweigh the loss of utility by any
other individual by virtue of its sheer magnitudes of utility gains. We can start to see how such an issue would not seem so outlandish in such a scenario when we remove certain constraints on possibilities. By removing these constraints, which I believe A.I. are poised to do uniquely compared to all beings up until now, we can create the spiral of destruction that a utility monster would cause. Thus, the dismissal of the problem of utility monsters as an impractical scenario isn’t one we can reasonably make without further justification.

There’s a further complication that makes this issue one that’s difficult to deal with from within the frame of a consequentialist theory. To nullify the potential issue of a utility monster, one may make the move with one’s theory to try and place some sort of conditional aspect upon the moral good to be maximized. For example, we can consider the maximization of utility if and only if there is some sort of distribution of utility that meets some prior condition or threshold. So an option might only be viable for our theory to consider if we have Smith and Jones having roughly equal end utility, or at least reasonably fair, rather than one having wildly greater utility than the other even if this inequality would actually maximize utility.

This has been a not uncommon move in the sophistication of Utilitarian theories (Neumann 1992), and is one that could, in theory, solve the issue by denying A.I. the possibility of taking all possible resources by virtue of their sheer utility gains. However, this presents a fundamental problem between the theories themselves and what they purport to do. If a theory is said to be consequentialist, it must be concerned with, namely, the consequences of each act. But if we adopt a principle that is prior or superior to utility, then we have the conundrum: “Is this theory really consequentialist then?” How can there be a moral principle prior to our primary moral good? If it is prior, then it seems like we are first and foremost dedicated to that principle,
and use utility as a tie breaker of sorts between possible options, which does not seem to reflect that we take utility as our moral good.

Another possibility one could argue from could be to ask what obligations we owe to A.I. who don’t even exist yet. For this argument, we turn to Partha Dasgupta (2004) to get a general sketch of this defense. For Dasgupta, there are 3 categorizations we consider populations under, and that is between present persons, which ought to be self explanatory, future persons, and potential persons. Future persons “aren't alive now, but will be alive in the future.”, and potential persons “are people who will be alive only if someone chooses to create them.” (Dasgupta 2004 p. 11). Obviously, our strong A.I. would be considered, at first, to all be potential persons; people who we have chosen to bring into the world.

Furthermore, under Dasgupta’s account, present persons and future persons have a stronger claim of obligation upon us than potential persons in a way that lets us override their considerations if need be. Dasgupta writes to this effect: “[An actual person] is a person; it is her well-being that is under consideration. She has a right to demand that she be acknowledged… Future people will have such claim rights, of course... It is potential people who cannot be said to have a corresponding right.” (Dasgupta 2004 p. 13) Future and actual persons, those will and do exist, have a stronger claim of obligation upon us than those potential persons who might exist.

If it is true that we have no or a lesser obligation to bring about potential persons, even if their existence in terms of strict consequences would be better than the alternatives, if it comes at the cost of present and future persons then we do not grant them a right to existence. From there, when we consider the potential persons in the form of A.I. versus the present and future persons whose lives would be demonstrably worse off with the creation of a utility monster, we deny that
it would be morally obligated or even acceptable to create such a being. If this is our position as a consequentialist, then we’re able to escape the problem altogether.

However, I have two responses to this. The first is to reiterate my objection that this defense is incoherent if consequences are taken to be our primary moral concern and good. Comparisons of utility take place in the context of possibilities. If utility can be taken to exist in some capacity in one choice, then we must count it. The distribution of utility we take to be important, such as the greatest average or the greatest total, is flexible. utility remains primary in that sense.

Dasgupta’s move is to try and prevent an accounting of potential person's utility in order to block the inference therefore that states of affairs where we bring about as many people as possible are considered morally good. However, doing so creates complications in the fact that this limitation is only placed on possible people. States of the world that could exist, if we brought them about through deliberate action. Future persons, who Dasgupta argues do have claim upon our actions now, will only exist if we make them exist, and turn them originally from potential to future persons.

This creates a problem, because future persons only exist by being potential persons initially, but what is precisely at stake is whether consequentialist arguments endorse making potential persons into future persons. Put more simply, what is it that allows some potential persons to become future persons and have claim over us, but not others when all potential persons could become future persons? If our question is whether or not bringing a person into the world would maximize utility, we cannot discount a potential person’s utility coherently, as we would have to discount every future person’s utility as well.
To make Dasgupta’s defense sufficient then, is that we have to assume a particular picture of how the future is. Our view of time has to have some belief that the future is “certain” in some respects but not in others. Dasgupta wants to allow for some maneuverability in this, given that potential persons are even possible in the first place. If the future was completely deterministic, for example, “potential people” would be impossible given that either they would exist, or they would not and we could not decide to bring them about. Given that choice is crucial in the distinction between future and potential persons, any view on time that states there is a fact of the matter in the future regarding potential persons is immediately ruled out. We cannot say that there is a fact that X will exist or not, or else they would not be merely a potential person. If it was true X would exist, then we could not stop it, and they would be a future person. If it was false X would exist, then we could not bring them about.

Since what differentiates future persons from potential persons is that potential persons can be chosen by us to be brought about, this creates a major issue. If we cannot distinguish between future and potential persons, then Dasgupta cannot exclude potential persons as not having moral claim upon us without also saying future persons have no claim on us. On the other hand, Dasgupta also believes that there is a definite group of persons in the form of future persons that will exist, which is how we can have obligations to them in particular.

This position on time is a non-trivial one, and one we may not want to commit to, but without which, the project falls apart. Consider having a child or not. There now exists a potential person for us to make into a future person. Suppose you then decide to have one. But now you think of having another afterwards, and you decide against it. What made that child different from the first one until the moment of your decision? Without Dasgupta’s assumption
of facts about the future, unless there is already some fact of the matter about a particular child’s existence, we could not tell.

If we took some less deterministic view of the future, this defense wouldn’t work. For example, if we had a more open view towards facts of the future, all future persons would be mere potential persons. If one believed that the future was completely indeterminate and there were no fact of the matter about it, then the concept of a “future person” would not be a tenable distinction. There would be no “fact” to check them against to sort them from potential persons. Under this account, future persons cannot have any claim upon us until they’re decidedly actual in some sense, so how can we have any obligation towards them, and how can their utility be distinct from other potential persons?

Worse still, without this deterministic assumption, we may wind up even arguing something entirely antithetical to Dasgupta’s position. If future and potential persons are on equal moral footing under this open account of the future, then that may even lead us to conclude that we have no obligation to future persons either! This view would go in the exact opposite direction we wanted in the first place, causing us to abandon concern for later generations all together. This would be disastrous for obvious reasons, so it becomes all the more vital that we buy into this assumption. If we do not want this view of time, however, then we have to look elsewhere for our defense.

So to even get this argument off the ground and before we can even consider reasons for and against adopting it, we already must buy into a particular view on time. This can put strains on commitments we may have elsewhere and thus may turn the whole project into a nonstarter. This is not necessarily a decisive problem on its own, but it is something to keep in mind when considering the arguments against it. With this all in mind, what are the arguments against this
position, for perhaps they may be weak enough to be overcome, and if we have no strong particular view on time, it may be worth adopting this argument. However, I believe that the drawbacks to hold this position are too great to solve the problem that Dasgupta wishes to solve.

For example, we cannot claim that some utility is lesser than another in our consideration so easily. The impartiality of the Utilitarian calculus has often been trumpeted as a driving reason for adopting the theory. The utility of the king is no more or less weighted than the utility of the peasant. If some forms of utility are more important than others, namely, future and present persons over that of potential persons, then it cannot be that all utility is equal in the eyes of the calculus. If we are to argue through this route, then we must give up on the claim that we are being wholly impartial about utility. So it is not at all obvious how Dasgupta’s argument can get us out of this predicament without losing something elsewhere.

The second problem is that future persons aren’t entirely a defined group. Compare to the case of present persons. It is a more or less a rigid group with relatively defined members. However, for the concept of future persons, we don’t have that luxury. Dasgupta uses the example of predictions of world population size in 2050. We have a rough idea that there will be 9.5 billion or so people alive at that time, but we have no idea who many of those people might be. Future persons only function as a collective group, where we know that they will exist and that we have obligations to them.

Given that future persons don’t magically appear from the aether, but rather are brought about by human choices, it must be the case that all future persons are potential persons until they become present persons. As we already established, if there is a heterogeneous mixture of humans and A.I., that can cause us to fail our obligations to future persons by creating this utility monster. This might even lead our hedonistic moral theory to actually endorse some sort of
radical view. One where the morally greatest world might even be one where humans do not exist any more. Given that a mixture of humans and A.I. cannot exist as utility maximizing, and that A.I. are utility monsters, they would be the superior choice.

This has a compelling justification for itself within a consequentialist theory. If A.I. are utility monsters, then obviously they must experience greater levels of utility as a whole than humans would. If we have no obligation to bring any potential person into existence, then we have no obligation to bring humans in particular into existence. So if we wish to maintain our obligation to future persons AND our commitment to A.I., then our answer is obviously to just entirely replace humanity with A.I..

If we populated the world exclusively with humans to avoid the utility monsters, then we’ve missed out on bringing about possibilities with better distributions of utility and have failed our primary moral theory. However, if we bring about humans and A.I., we’ve held to our moral theory, but due to their great suffering, we’ve failed many future persons to misery and failed the obligations associated with them from Dasgupta’s account. With a pure A.I. population, however, we avoid either issue, given that all of them experience great levels of utility, and none of them are condemned to suffer at the hands of utility monsters.

This obviously is an undesirable and repugnant conclusion to most anyone except the most radical of anti-natalists, and in fact, it gets us in the opposite direction of what we’re looking for if we wish to have humanity continue into the future. Dasgupta’s concepts of potential and future persons, due to their inherent inspecificity, cannot work as a defense against this conclusion then. This leaves us with only one real direction we can argue from, and it is one not wholly unfamiliar to consequentialists. Our answer to dealing with this potential utility monster has to be on empirical grounds.
If we cannot prevent A.I. from having a claim upon us to bring them about, then we have to reshift our focus to refute the claim that A.I. would be utility monsters from the physical facts. A potential argument might say that computational limitations might limit the ability of A.I. to feel enough utility to ever become utility monsters. I will address this argument later on in particular, but it is rough form of the kind of argument that we now need to handle. Rather than changing the Consequentialist theory in response to these objections, one might argue that the facts cannot or would not support the existence of utility monsters. Thus, we must also be able to demonstrate that it is indeed a live possibility such a world would exist in order to fully defend the position of this paper.

First, we need to clarify what is meant by “cannot or would not support the existence of utility monsters”. In this sense, I believe what we must deal with, specifically, is either the physical impossibility of it being able to happen (the laws of physics would not permit it in some capacity), or that it would be wildly impractical (it would be unreasonable to assume it could be produced on any significant scale.) Of these two, the former is the stronger claim, and must be shown to be false before we can begin to tackle the latter. Even granting this stronger claim, however, may not be a sufficient defence. There’s an additional angle this problem can approach us from. We can present the issue in a different way that doesn’t even require us to imagine that A.I. can hit such incredible high levels of utilities, but rather more human-possible degrees. All it takes to accomplish this is a reframing of how we conceive of a “utility monster”.

Rather than thinking of it as one particular entity, we can capture the same idea through a community who possesses certain similar characteristics in their members. While at the individual level they may pose no dire threat, in this collective sense, they gain a moral claim upon us. This not only allows for a larger scale effect in decision making in things like resource
distribution but can also reduce the “barrier to entry” in order to warp ethical decision making in the member’s favor without having to have particular members be so unrealistically hedonistic.

Imagine that fruit flies experienced utility on the same level of humans, so that their desires had to be weighted on the same level of a human when performing our moral calculus. In such a case, given that they experience utility roughly on par with a human, then their sheer numbers and easy satisfaction of desires would say that, in order to make the greatest good, we would have to allocate many, if not most of our resources to pleasing fruit flies. They require so little to satisfy their desires for things like food or water, whereas comparatively, we require so many to accomplish the same thing, so to most efficiently allocate resources, we would dedicate much of our efforts for their sakes.

It does not matter that the fruit flies do not experience extremely unbelievable levels of utility from our actions. Rather, the problem presents itself through a combination of two factors. Sheer numerical advantage, combined with the easy satisfaction of desire, causes the problem to return. Considered as a bloc, the fruit flies are the utility monster, because it becomes a trivial task to satisfy each fly’s particular desire, and since there are easily thousands of flies we could satisfy with the same resources it would take to sustain a single human, it becomes immediately obvious how we would be required to distribute resources. Instead of satisfying the one human, we could generate many times over more total utility, or a higher average utility, etc. from catering to these humanistic flies, especially when producing so many flies is an easy thing to accomplish.

However, it can be argued that if this would be true, then we’d already have run into this problem with humans alone. If this were the case, then it seems like there would be an obligation to generate as many children as possible so long as their happiness was greater than their misery.
We do not believe this is the case, so we shouldn’t necessarily believe that we have to do the same here with fruit flies or A.I. This is the issue that Narveson (1967) tackles, and I will appropriate part of it here for the sake of argument.

Narveson’s main thrust in his paper is that we have no obligation to bring persons into existence under a utilitarian account. This claim comes from his account of non-existent persons, which we will not touch on here. However, part of that justification also comes from the fact that he argues that with the addition of every new person, we’re shifting the calculus around with numbers in a way that doesn’t actually affect anyone in a way that “changes” utility in a meaningful sense.

For example, my existence does not affect the happiness of someone in Laos, except in the most strained sense. Likewise, the new existence of a person does not really affect the utility of those already existing for the most part. The primary benefit in our calculus of the existence of a person is that that person is now a new source of utility. Otherwise, we’ve not changed the world itself, like we would have with something like with providing famine relief, treating the sick, or even more mundane things like giving a friend a ride when their car is in the shop. Those actions affect those who exist, which is why they generate an obligation in a way that bringing a person into the world doesn’t.

To help justify this, Narveson creates a thought experiment for us to imagine. The example used is about a king of Fervians, where the king describes the fact that a group of very happy martians have become part of Fervian kingdom and that this will make the Fervian people happier through their addition to the kingdom. In a sense, the king is not telling a lie. The Fervians, as measured through utility, are objectively better off, as now their overall utility is much greater in aggregate and in terms of averages.
However, the lives of the Fervian people as they are now, are wholly unaffected by this acquisition. The average happiness of the Fervians went up because of the happiness of the martians being factored in, but Narveson argues, this seems like a bait and switch to the people who expected *their* happiness to increase. The king equivocates on the group “Fervians” in a sense as no-one was made happier, but rather, just changed around groupings of utility in the comparison between the Fervians of now, and the Fervians after the acquisition to include the happy martians.

This case is analogous to the case of having children, or in our case, A.I. who are happier and thus would be required to bring about as much as possible. None of the people who exist are improved by the addition of happier children or A.I. (except by “indirect means”, as Narveson puts it, which describes incidental factors from a person’s existence), so we’re performing a sleight of hand in a way by saying that the world is in a better state by bringing happier persons into it. And if that is the case, then we can resist the conclusion of an obligation to bring about A.I. by claiming utility hasn’t actually changed in any meaningful sense, we’ve just shifted it around a bit.

I believe this analogy fails, however. Let’s reexamine the case given to us. We have the happy martians, and the Fervians, and the combination of the two groups producing a “better” outcome. So far, yes, we’ve performed a sleight of hand that Narveson claims it is. But I do not believe that this is analogous to bringing a child or an A.I. into the world. The Martians did not just spontaneously appear in the world. They already existed and were accounted for by our calculus. So yes, combining them into the Fervians did nothing to change utility for anyone, because none of the facts relevant to changing utility were altered.
But this is not the case for bringing a new being into the world. In that situation, we’ve added or changed the number of individuals involved in our total considerations that can’t be accomplished by just moving them around. So we’ve done more than just shifted the numbers around in bringing an A.I. or a child into the world, we’ve fundamentally shifted how much pain or pleasure is in the universe, in addition to how many people we have to think about.

To use my own analogy, Naverson’s example is like an hourglass filled with a certain amount of sand. Turning over the hourglass moves the sand around to different places, but it doesn’t change how much time it measures, just where the sand is distributed. Bringing a new person or A.I. into the world, however, is like adding grains of sand to the hourglass. It changes the rate at which it falls, how long it takes to filter to the opposite end, etc. How the hourglass works in a way through its myriad of interacts by its mere introduction. Thus, numerical changes in our considerations of utility matter in a way that goes beyond mere shuffling of utility and cannot be ignored.

If this is the case then it doesn’t seem immediately obvious that Naverson’s defense is tenable. Changes in numbers matter in significant ways to our calculations of utility that cannot be captured in a similar way just by moving around already existing agents in the world. So we are obligated, in some sense, to consider how utility functions respond to the possible addition of agents like A.I., and that is all we need in order to cause our explosion of utility. For if we could improve the world by the addition of one robot, we have no grounds to say no to the next and the one after that, and so on, until it would no longer be desirable in terms of utility to add one more.

Now it would be fair to ask that if my conclusion is true, this would also apply to humans and that consequentialists would be forced to endorse the conclusion we ought to bring about as many humans as possible until we hit some sort of Malthusian state where the addition of one
more human would not increase utility, but rather, decrease it. However, consequentialists can still reject this position. I will touch on how one can go about this for humans, but not A.I. later on. But if that is true, then one might ask, what has all this been for? The answer to that question is that this rejection must be based on our old friend, empirical reasons.

If this argument and the one regarding constraints on utility are true, then it does not appear to be the case that, a priori, consequentialists can coherently reject creating A.I. on a mass scale, except by reason that the facts in such a case would not find that situation conducive to utility. Nothing would be able to hold back consequentialists from being coerced into endorsing this fruit fly world, except for the fact that fruit flies don’t actually experience utility as such. This means that my route of argumentation is clear: I must demonstrate that the existence of A.I. would be utility maximizing by virtue of being utility monsters due to various reasonable empirical facts. If I am able to do this, then it would entail a “moral superiority” of A.I. under a consequentialist account. That in turn would require us to relinquish our claims upon anything to better satisfy A.I., which seems undesirable from practically any point of view that deeply values our own interests in any sense.

With that all in mind, if we have the assumption that strong A.I. would experience utility on a level roughly to our own, combined with two quite reasonable assumptions, we have a recipe for disaster. The first assumption is that the major difficulty in creating A.I. is just creating the very first one. However, once we figure out how to create them, the production of every A.I. afterwards becomes a lot easier. It seems like all you would have to do would be to replicate everything that went into the creation of the first A.I., from the frame to the programming that directs it. This means that A.I. could be reasonably produced at scale allowing for great numbers to be created with ease. This sets up part of the issue, but not entirely. After all, if it took a heavy
investment of resources to produce or sustain each individual unit, then our resources will be depleted quite quickly, and so the calculus would start to impose limits on how many we could produce.

The second assumption then, is that A.I. would not require as heavy a resource investments as humans in order to satisfy their desires and to sustain them. We need food, water, shelter, warmth, personal space, medicine when we’re sick, time to relax and play, the list goes on. However, what would an A.I. need to maintain itself? Power, spare parts, a mechanic, and programmer, and it seems like not much else unless we gave them various predilections for other things. With our finite resources, it seems far easier to meet the needs of an A.I. than it does a human, and so it seems like we ought, for the sake of efficiency, to dedicate our efforts to satisfying A.I. instead of ourselves. After all, why please ten humans when you could please twenty A.I. to the same degree with the same effort?

So now we can see the problem presents itself in a much more direct way. In a world with scarcity of resources, there is a decision to be had about how to allocate such resources and where to dedicate our efforts in order to generate such resources. If our aim is to bring the greatest joy to the greatest number for example, then the conclusion seems inescapable: we humans ought to give all the resources we can into producing as many A.I. as possible and give them the lion’s share of resources if not everything altogether, for they can be satisfied with much less than we can, and in much greater numbers.

Additionally, this objection can extend to other forms on consequentialism. If we can satisfy the same number of desiring agents with less resources, then with more resources now freed up, it follows that we have those resources to focus on satisfying other desires, causing the average level of satisfaction or utility to rise, as more desires are able to be catered to. So it’s not
simply that this argument applies to aggregate or maximizing theories on consequentialism, but all potential distributions of our consequentialist good.

To make another analogy, we have the choice of refueling two cars. One needs much more maintenance, is a bit more finicky in how it runs, and is noticeably less fuel efficient. If our goal is to get from A to B, (satisfy desires), it seems obvious that the one that is less of a hassle and inefficient would be the one we would choose to use. It wouldn’t matter if we were trying to go back and forth between A and B as much as possible, or trying to get from A to B as easily as possible, etc. By all reasonable measures, if we’re trying to accomplish our goal, then the more efficient and consistent option is the one we would choose to do so. And so, it seems like, for any given decision in bring about utility, all decisions would inherently favor satisfying A.I. rather than humans, possibly to the point of requiring ourselves to sacrifice everything to please more A.I.

An objection might be raised, however, that running a strong A.I. might be an immensely difficult task, especially given modern computing power as we know it. The sheer need for power and energy would be disastrous in terms of sustainability and would make it hard to continually satisfy great numbers of A.I. at once. A comparison might be made to something like cryptocurrency mining, where huge amounts of computational power is thrown at number crunching in a way that is incredibly inefficient in terms of power to profit except at large scale. To this objection, I have 2 counterarguments.

Things like cryptocurrency mining are intentionally designed to be incredibly inefficient to do. The resources required to produce them in any significant amounts is astronomically high in order to maintain the value of the currency. If the currency was incredibly easy to produce at scale, then there would be nothing to support the worth of the currency. Conversely however,
attempts to make strong A.I. are intentionally going to be made more and more efficient in order to make strong A.I. more practical. So while the intensity of resources required for something like Bitcoin goes up exponentially, we would see the reverse trend for A.I., allowing them to be more feasible as time goes on.

Second, there’s no reason to believe that current computational limits are indicative of the kind of power we will have at scale in the future. We may yet have another major computing revolution, allowing us to have computers magnitudes and magnitudes more efficient and powerful than what we have now or could even conceive of. Even if running a strong A.I. took fundamentally high magnitudes of computational power, we answer that with faster, better, and more efficient computers, improved cooling techniques, etc., allowing for strong A.I. to reasonably exist and survive without excessive demands, which would only continue to diminish with more innovation and efficiency increases.

The other side of the coin, however, is that this objection is focused most squarely on the hedonistic forms of consequentialist theories. A theory that advocates the maximizing of something like aesthetic beauty, or bonds of trust, or friendship between rational creatures isn’t necessarily going to feel the full force of this problem. I do believe there are in principle ways we can extend this argument in order to target those kinds of theories as well. For now, however, given that hedonistic forms of consequentialism are very popular and are often the most intuitive, I do not think that takes away from the power of this problem. If a sufficient number of the mainstream views of consequentialists are affected, I think that constitutes a large enough problem to need dealing with in order to maintain one’s view if one is committed to applying a theory to the ethical treatment of A.I.
Section 4: Prospects for a Better Theory:

So, the question may be asked, is there any hope for the salvaging of a consequentialist theory for A.I.? After all, if my arguments are taken to be accurate, it does not appear that there is a clear answer to how we can go about even beginning to prevent these conclusions. I believe the prospects are grim to say the least. However, there are a great deal of many clever philosophers, who can create even more clever positions than I may and be able to overcome these arguments or their conclusions through their ingenuity. So I will not say that there is no possibility of recovery. But I will say that it is not obvious how one can escape these conclusions.

So for now, let us take these conclusions as inescapable. Let us take stock of our options, and there are three. We can accept our replacement by A.I. as morally required, but I believe this is a non-starter for obvious reasons. Instead, we can abandon the project altogether and say something to the effect that ethical treatment of A.I. is impossible as no theory is adequate for them, given that their nature precludes the only acceptable possibility. However, I do not believe this is an option for many. If we were already committed to finding a theory that would guide our action towards A.I., then simply losing one prospect does not seem to merit relinquishing our quest. To use a Poker analogy, many people are not going to go all in on one hand only to fold before the river. We want to see that every possibility is used up before we admit defeat.

This brings us to our third option, and that is simply to use another type of theory. Instead of using a consequentialist one, we could employ a deontological theory, for example, to capture and give structure to our moral obligations. Of course, as we saw with consequentialist theories, this is not always as straightforward a matter as it might appear. So we need some assurance that this will not fall apart too, lest we get our hopes up too hastily. Thus, in this section, I will be
using a vaguely Kantian-esque frame of reference. This kind of theory carries with it its own set of potential problems (Tonkens 2009, Coeckelbergh 2010, Basl 2013). However, the aim for now is simply to show how these current objections can be better met by a different ethical theory.

This is not to commit myself to a Kantian theory in particular. Rather, I employ it as a model, to show how a deontological system may maneuver around the problems we saw before. Given that many deontologists use much of Kant’s framework as the basis of their own to expand off of, similar responses should be able to be provided as a result. I believe even non-deontological theories may make similar motions, but obviously, I cannot cover every conceivable example, so I must limit my scope for now to only give deontological answers.

The objection we must deal with then, is that there would become an obligation to create a feedback loop. There is no obligation to create incredibly happy A.I., nor is there an obligation to maximize their numbers. Of course, there is a weak obligation to make A.I. happy, like we would consider with humans; we’d have a similar obligation to them like we might have towards our own children (Schwitzgebel and Garza 2015). But, using the universalization formulation, we can see how our obligations for the well-being of our fellows in the kingdom of ends can only go so far. So, despite A.I. even potentially being incredibly happy at receiving a particular resource, far greater than any human could hope to feel, that does not produce within us an obligation to therefore give everything to them instead of keeping some for our own ends.

In this way, we can resist the explosive conclusion out the gate, given that we are freed of the fundamental problem. The utility monster is no problem for us. However, obviously, this may seem like sophistry. Of course utility monsters are no problem for non-consequentialist theories, so it may appear like I have accomplished nothing in arguing this point. However, I
point to this fact of the Kantian theory in service to answering the problem that might seem more relevant. I wish to highlight that within a system of rules, there are means of negating the “forcing” of obligation within us in service to another, regardless of the facts of the world. It does not matter that I stand to gain more than you do for a promotion or even that I may do a better job doing it. I am not morally entitled to that promotion.

This brings me to the equivalent problem for a Kantian theory. Suppose that the A.I. created become far more intelligent and rational than even the smartest of humans. Would they then have greater moral weight in ethical evaluation? It may be analogous to comparing a human life to that of an animal’s. We obviously ascribe greater moral weight to ourselves, given that we have rational faculties, so by further comparison, would we stand in that relation as animals compared to A.I. should they become hyper intelligent? Would they be able to make greater claim upon resources by the fact that they could use it better, or that they are more deserving by virtue of being far more rational than us?

The answer here is no. There is no greater obligation generated by virtue of being more rational, but rather by merely having rational faculties. Einstein does not have a greater claim than me regarding any matter as far as morality is concerned, we are equals, no matter how much more intelligent he may be. Increasing this divide further in the case of our incredibly intelligent A.I. makes no difference whatsoever. We as members of the kingdom of ends all share the same considerations when it comes to matters of morality. As such, we are under no obligation to forgo our interests to further the interests of another. The analogous problem to utility monsters in Kantianism is defeated and need not concern us further.
Conclusion:

A fair question to ask after all this “Where does this leave us?” The answer to that question, at the very least, should be that it is non-obvious that we can use a utility-centric ethical theory in regards to actions towards A.I. without some significant potential problems to deal with. I leave room that there may be some sort of way these theories can be salvaged, but it is not apparent to me how. Rather, I believe it may be more productive to look elsewhere for an ethical theory if we wish to treat robots in a moral manner. Many deontological and virtue ethics based approaches, among many others, remain quite tenable as section 4 demonstrated. These theories are more likely to be fruitful and be the path of least resistance, which gives us all the more reason to adopt them instead.
Works Cited:


