Routledge
Taylor & Francis Group

Ethics, Policy & Environment

ISSN: 2155-0085 (Print) 2155-0093 (Online) Journal homepage: www.tandfonline.com/journals/cepe21

The Rebugnant Conclusion: Utilitarianism, Insects, Microbes, and AI Systems

Jeff Sebo

To cite this article: Jeff Sebo (2023) The Rebugnant Conclusion: Utilitarianism, Insects, Microbes, and AI Systems, Ethics, Policy & Environment, 26:2, 249-264, DOI: 10.1080/21550085.2023.2200724

To link to this article: https://doi.org/10.1080/21550085.2023.2200724

Published online: 26 Apr 2023.

ETHICS, POLICY & ENVIRONMENT
2023, VOL. 26, NO. 2, 249–264
https://doi.org/10.1080/21550085.2023.2200724

Routledge
Taylor & Francis Group

The Rebugnant Conclusion: Utilitarianism, Insects, Microbes, and AI Systems

Jeff Sebo

Department of Environmental Studies, New York University, New York, NY, USA

ABSTRACT

This paper considers questions that small animals and AI systems raise for utilitarianism. Specifically, if these beings have more welfare than humans and other large animals, then utilitarianism implies that we should prioritize them, all else equal. This could lead to a 'rebugnant conclusion', according to which we should, say, create large populations of small animals rather than small populations of large animals. It could also lead to a 'Pascal's bugging', according to which we should, say, prioritize large populations of small animals even if they have a low chance of being sentient. I suggest that utilitarians should accept these implications in theory but might be able to avoid some of them in practice.

ARTICLE HISTORY
Received 27 March 2023
Accepted 5 April 2023

KEYWORDS
Utilitarianism; animal ethics; AI ethics; creation ethics; population ethics; duties to future generations

1. Introduction

We are currently in the midst of rapid moral circle expansion. Animal advocates have made significant progress over the past fifty years by promoting the idea that we have moral duties to domesticated animals. We are now in the early stages of promoting the idea that we have moral duties to wild animals. Some of us accept that we have such duties because we think that we should help others when we can. Others of us accept that we have such duties because we think that we are harming many of these animals, and that we should reduce and repair these harms when we can. Regardless, the idea that we have duties to many nonhuman animals is fast gaining acceptance (Johannsen, 2021; Palmer, 2010; Sebo, 2022).[1]

This moral circle expansion raises many difficult questions about our moral priorities. For instance, humans are currently harming and killing tens of billions of domesticated animals per year and hundreds of billions of wild animals per year (Sanders, 2020). At least in terms of scale and neglectedness, then, our duties to current and near future nonhuman animals would seem to take priority over our duties to current and near future humans, all else equal. Granted, we might think that we should prioritize humans for other reasons, including reasons involving tractability and indirect effects, as we will see. Still, we are slowly coming to terms with the idea that nonhumans matter much more than we previously thought.

CONTACT Jeff Sebo jrs477@nyu.edu Department of Environmental Studies, New York University, 285 Mercer Street 8th Floor, New York, NY 10003, USA

But as substantial as this moral circle expansion has been, it is not nearly complete. For instance, our discussion of duties to captive and domesticated animals tends to focus on animals such as cats, dogs, cows, pigs, and chickens. And our discussion of duties to free and wild animals tends to focus on animals such as chimpanzees, elephants, koalas, dolphins, and polar bears. While there is no single category that includes all these animals, in general we seem to focus more on large animals than on small animals, on vertebrates than on invertebrates, and on land animals than on aquatic animals. The result is a moral community that is many times larger than it was before, but still many times smaller than, I believe, it should be.

When we take seriously the possibility of a moral community that includes all sentient beings – large and small, vertebrate and invertebrate, terrestrial and aquatic – we realize that this next expansion might, if anything, be even more transformative than the last one. The world is full of conflicting interests and needs, and it is also full of very different kinds of populations. For instance, if we have to choose between improving the lives of a small number of large animals and improving the lives of a large number of small animals, then which should we choose and why? This kind of case requires us to think not only about what we owe each kind of animal but also about how to compare these duties when they conflict.

Suppose that we determine that large animals like humans have more welfare on average but that small animals like insects have more welfare in total. What follows for ethics and politics? Which populations should we prioritize within each generation, all else equal? And which populations should we prioritize across generations, all else equal? Suppose further that we determine that many beings, including microscopic organisms and current and near future AI systems, are at least *possibly* sentient, and that the size of these populations relative to insects rivals the size of insect populations relative to humans. How, if at all, should that possibility affect our moral priorities within and across generations?

My aim in this paper is to survey these questions from a utilitarian perspective, building on work from Horta (2010), Ng (1995), Tomasik (2015), and others. I will show that utilitarianism implies that insects can take priority over humans and that microbes or AI systems can likewise take priority over insects. Granted, we might still have reason to prioritize humans at present and in the near future, since our duties regarding the distant future outweigh our duties regarding the present and near future, and since improving human lives at present and in the near future is key to improving the distant future. But in this case, what saves utilitarians from one surprising conclusion might be another surprising conclusion.

To be clear, while I focus on utilitarianism here for the sake of simplicity, I think that other moral theories face versions of these questions as well. Any theory that involves a duty of beneficence or a duty of non-maleficence will have to deal with questions about, say, how to set priorities between small populations of large animals and large populations of small animals, since there might be many cases where we have the power to help or avoid harming *either* the former populations *or* the latter populations but not both at the same time. So while my discussion here might focus on how one moral theory might navigate this strange future, we should keep in mind that this strange future awaits us all.

2. Background

Any discussion about how to set priorities must begin with a certain set of normative and empirical assumptions. So I begin in this section by clearly stating the assumptions that will govern my discussion here. Normatively, I will assume a classical utilitarian theory of morality, according to which we morally ought to maximize positive welfare in the world.[2] And empirically, I will assume that many beings might have the capacity for welfare, that beings with larger brains and longer lifespans tend to have a higher capacity for welfare than beings with smaller brains and shorter life spans on average, and that many other factors will be relevant to our priorities too, such as tractability and indirect effects.

Consider first my normative assumptions.[3] Utilitarianism, as I will be interpreting it here, involves a *hedonist* theory of the good. According to hedonism, the only intrinsically and finally good state is pleasure, that is, positively valenced conscious experience. And the only intrinsically and finally bad state is pain, that is, negatively valenced conscious experience. Everything else is only, at most, extrinsically or instrumentally valuable, depending on its relationship with pleasure or pain. For example, to the degree that knowledge promotes more pleasure than pain, knowledge is good according to hedonism. But to the degree that knowledge has the opposite effect, knowledge is bad according to hedonism.[4]

Utilitarianism, as I will be interpreting it here, also involves a *totalist* theory of the good, according to which the world is better when it contains more pleasure and less pain in total. For example, if one world contains more pleasure in total and another contains more pleasure for the worst-off, utilitarianism implies that the former world is better. Of course, we might still have reason to prioritize the worst-off in many cases on this view, since, for instance, an additional $1,000 will tend to make a bigger difference for low-income person than for a high-income person and, so, distributing this money to the low-income person will tend to do more good, all else equal. Either way, what does the most good is best.

Finally, utilitarianism, as I will be interpreting it here, involves a *maximizing act consequentialist* theory of the right. According to act consequentialism, the rightness or wrongness of an action depends on the consequences of that action. And according to maximizing act consequentialism, an action is right if and only if it maximizes the good (or, in the scalar formulation, an action is right to the degree that it maximizes the good). And when we combine maximizing act consequentialism with hedonism and totalism, we get classical utilitarianism, as I will be interpreting it here: An action is right if and only if, or to the degree that, it produces the most pleasure and least pain possible, in total.

At least in theory, utilitarianism is highly demanding. It implies that we might sometimes be required to sacrifice our own well-being for the sake of the greater good or sacrifice the few for the sake of the many. Granted, we might think that there are limits to how demanding utilitarianism can be in practice, since, for instance, we need to take care of ourselves to be able to take care of others, and we need to treat others as having rights to maintain the kinds of norms and structures that allow us to maximize pleasure in the long run. But these considerations might not always be decisive, and even when they are, utilitarianism might still be more demanding than commonsense morality in many cases in practice.[5]

As this last point suggests, I am not assuming utilitarianism here because I reject other moral theories. Far from it. As I argue elsewhere (Sebo, 2022), I think that most people in most situations should accept a pluralistic, partly consequentialist and partly non-consequentialist moral theory in practice. For instance, utilitarians should accept such a theory since we sometimes need to respect rights to promote welfare, and rights theorists should accept such a theory because we sometimes need to promote welfare to respect rights. We should also accept such a theory on strategic grounds, since it can serve as the basis for collaboration between consequentialists and non-consequentialists in advocacy and policy.

Instead, I am assuming utilitarianism here simply because the issues that I discuss in this paper are complex, and it would be difficult to assess them from multiple moral theories at once. So my strategy will be to (start to) assess these issues from a utilitarian perspective for the sake of simplicity and specificity. My hope is that we can then assess these issues from other perspectives as well, and that we can then put these analyses together to arrive at a preliminary view about how to set priorities in a multi-species community that includes small populations of large animals and large populations of small animals (and other possibly sentient beings). To that extent, my analysis here will be highly preliminary and incomplete.

Now consider my empirical assumptions. My first empirical assumption is that many beings at least *might* be sentient (that is, might have the capacity for welfare), given the evidence available. Granted, we might think that some beings are more likely to be sentient than others. For instance, we might think that mammals are more likely to be sentient than insects, and that insects are more likely to be sentient than microbes or current AI systems. But we are not, at present, able to rule out the possibility that any of these beings are sentient. Even if we assign an *astronomically low* probability to the idea that, say, microbes are sentient, we should not assign a probability of *zero* to that idea at present.

My reason for making this assumption is that the problem of other minds limits how much we can know about other minds at present and for the foreseeable future. Since the only mind that I can directly access is my own, I am not able to directly confirm or disconfirm what, if anything, it might be like to be anyone or anything other than myself. And while some theories of consciousness, such as higher order thought theories, imply that only some of the beings listed above are conscious, other theories, such as panpsychist theories, imply that they all are. Unless and until we make substantial progress on this problem, we will need to keep an open mind about the scope of sentience in the world (Carruthers, 2004; Sebo, 2018).

My second empirical assumption, which I gestured at a moment ago, is that beings with larger brains and longer life spans will tend to have a higher capacity for welfare than beings with smaller brains and shorter life spans. For instance, we might think that an individual human can experience more pleasure and pain than an individual insect, and that an individual insect can experience more pleasure and pain than an individual microscopic organism (assuming, of course, that these beings can experience any pleasure and pain at all). Granted, we might think that some of these beings experience *astronomically little* pleasure and pain, if any at all. But if they are sentient, then they might experience at least *some*.

My reason for making this empirical assumption is that it seems plausible that an individual with more complex neural systems related to pleasure and pain would be capable of a higher amount of pleasure and pain at a time, all else equal. It also seems plausible that an individual with a longer life span would be capable of a higher amount of pleasure and pain over time, all else equal. Granted, there are many complications here, and it is unlikely that the capacity for welfare at a time will be a simple function of neuron counts, or that capacity for welfare over time will be a simple function of life spans (Schukraft, 2020). But this general assumption is all I need to motivate the questions that I will ask in this paper.

My third and final empirical assumption is that other factors, such as tractability and indirect effects, are relevant to which beings we should prioritize according to utilitarianism. First, we need to consider the tractability of the problems that we face. For instance, even if our actions impact many insects, we might not have a duty to help these insects in practice if we have *no knowledge at all* about how to help them or *no power at all* to help them in particular contexts. I will suggest that we should, in fact, limit how much we prioritize insects and other such beings at present for these reasons. But I will also suggest that we should still prioritize these beings much more than we do, and much more in the near future than at present.

Second, we need to consider the indirect effects of our solutions to these problems. In particular, even if our impacts on near future insects matter more than our impacts on near future humans *all else equal*, our impacts on near future humans might matter more than our impacts on near future insects *all things considered*, since what happens to humanity will likely determine how many sentient beings can exist and how good their lives can be in the long run. I will once again suggest that we should, in fact, limit how much we prioritize insects and other such beings for these reasons. But I will also, once again, suggest that we should still prioritize them much more than we do, and much more in the near future than at present.

If we put these normative and empirical assumptions together, then we find that utilitarianism might be an even more radical departure from commonly accepted moral norms than we expected. At least in principle, it implies that we can have a duty to prioritize insects over humans and microbes and AI systems over insects. And while utilitarians might be able to avoid some of these implications in practice given the numbers involved or considerations involving tractability and indirect effects, we will likely not be able to avoid *all* of them in these ways. No matter what, a commitment to doing the most good possible will likely require us to prioritize beings who are very unlike us, either in the short term or in the long run.

3. The Rebugnant Conclusion

Utilitarianism might have surprising implications about our duties regarding future insect populations. Plausibly, maximizing utility in the future partly requires creating a world in which the global population can flourish in relative harmony. But it also partly requires considering which beings to add to this global population in the first place. For instance, should we aim to bring about a world with a higher ratio of large to small animals, a higher ratio of small to large animals, or a balance between the two? While asking such questions might seem like playing God, the reality is that humans are already shaping

future populations whether we like it or not. We need to be thoughtful about how we wield this power.

This question is related to a problem in population ethics that Derek Parfit calls *the repugnant conclusion* (Parfit, 1984, pp. 381–90). Here is the problem. Suppose that we have to choose between two future populations: Population A contains one million people with one million units of pleasure each, and population B contains two million people with 999,999 units of pleasure each. Which one should we choose? Many people have the intuition that we should choose B, since there are twice as many people and everyone is nearly as happy on average, and there is nearly twice as much pleasure overall. This intuition is friendly to utilitarianism, which implies that we should maximize total utility, not average utility.

So far so good. But now suppose that we take this reasoning farther. Instead of choosing two million people with 999,999 units of pleasure each, we choose four million with 999,998 each. Then we choose eight million with 999,997 each. Eventually we choose a population where everyone has only one unit of pleasure, but where the population is so large that they still have nearly twice as much pleasure as the previous one overall. Many people have the intuition that we should *not* choose such a population; in fact, Parfit goes so far as to call the idea that we should choose this population *the repugnant conclusion*. This intuition – which, to be fair, not everyone shares (see, for instance, Ng, 1989) – is less friendly to utilitarianism.

While many people default to thinking about the repugnant conclusion in same-species cases, we can think about it in multispecies cases as well. In a same-species case, we might compare small human populations with high average welfare with large human populations with low (but still positive) average welfare. In a multispecies case, we might compare small human populations with high average welfare with large, say, ant populations with low (but still positive) average welfare. Either way, the outcome is the same in both cases: One population contains much more pleasure on average, the other population contains much more pleasure in total, and we need to determine which state of affairs is better all else equal.

Granted, there are many relevant differences between same-species and multispecies cases, particularly in the real world. For instance, we might wonder whether we have to choose between humans and ants at all and whether ants really do experience more pleasure than humans in total. But while these details might allow us to avoid this problem in some cases in practice, they do not allow us to avoid the problem in principle, or even, necessarily, in all cases in practice. *If* we have to choose between humans and ants, and *if* the ants would experience more pleasure than humans in total, *then* we should bring about the ant population all else equal, according to this view. We can call this implication *the rebugnant conclusion*.[6]

In general, when a seemingly plausible argument leads to a seemingly implausible conclusion, we have two options. First, we can accept the conclusion. For instance, in this case we can accept that we might sometimes be required to produce large populations of small animals rather than small populations of large animals, provided that doing so will, in fact, maximize utility. Granted, we might experience this implication as implausible. But all moral theories have *some* implications that we might experience as implausible. And given our self-interest, speciesism, status quo bias, scope insensitivity, and so on, we

should expect to experience some implications as implausible even if they are in fact correct.

Our second option is to reject features of utilitarianism that lead to this conclusion. For example, if we reject hedonism, then we might be able to deny that human and ant well-being are comparable (see, for instance, Korsgaard, 2018). If we reject totalism, then we might be able to deny that the world with the most pleasure in total is best.[7] And if we reject maximization, then we might be able to deny that we are morally required to produce the best world.[8] Philosophers commonly make such moves to reject the repugnant conclusion, and they can do the same here. My own view is that accepting this conclusion is better than rejecting some or all of these features of utilitarianism, but others might disagree.[9]

But even if we accept this conclusion in principle, we might be able to reject or constrain it in many cases in practice. First, we might worry about tractability. It seems relatively tractable to create more humans and other large animals *and* ensure that they have net positive experiences in the future. In contrast, while it seems similarly tractable to create more insects and other small animals in the future, it seems less tractable to ensure that they have net positive experiences, particularly insofar as they continue to reproduce through r-selection, that is, the evolutionary strategy that involves having a relatively large number of babies with a relatively high infant mortality rate (Groff & Ng, 2019; Horta, 2010; Tomasik, 2015).

Second, we might worry about the indirect effects of prioritizing large populations of small animals for the far future. That is, even if we feel confident that increasing the ratio of small to large animals would maximize utility in the short term, we might not be confident that it will do so in the long run. The reason is that if we want to maximize utility in the far future, then we might need to ensure that humanity can survive long enough to ensure that many beings will exist in the far future and have net positive welfare. This consideration might save us from a surprising neartermist implication, but it would do so only by replacing it with a surprising longtermist implication, as we will discuss in the next two sections.

Finally, we might worry about empirical facts regarding insects and other small animals. For us to face the rebugnant conclusion in the real world, it would need to be the case that we face a choice between bringing about a small population of large animals and bringing about a large population of small animals, that the population of small animals would have less welfare on average and more welfare in total than the population of large animals, and that this population of small animals would have net positive welfare rather than net negative welfare. If we reject any of these assumptions, then we might be able to avoid the rebugnant conclusion, at least in its pure form, in the real world in the near future.

My own view is that the first two points, about tractability and indirect effects, are plausible but not decisive. It seems clear that we can more effectively improve the lives of small populations of large animals than improve the lives of large populations of small animals at present, due to our epistemic and practical limitations. But this situation can change. The more we improve our knowledge about small animals and our power to help them, the more we might be able to improve their lives effectively at scale as well. And if and when that happens, utilitarians will have to take more seriously the possibility that we

should prioritize expanding and assisting these populations on grounds of scale and tractability.

I also think that the third point, about empirical facts regarding these animals, is plausible in this context. At least for now, all realistic options will involve bringing about mixed populations of large and small animals with good and bad lives. But as we will see in §5, we might face a choice between increasing or decreasing the ratio of large to small animals in the world. And whether or not small animals have net positive welfare, we should still consider them when deciding what to do. Specifically, if they have net positive welfare, then we should treat any policy that creates more or fewer of them as good or bad, respectively, to that degree. And if they have net negative welfare, then we should do the reverse.

With all that said, my guess is that at the end of the day, humans will not be required to prioritize bringing about insects and other small animals in the near future for reasons involving tractability, indirect effects, and uncertainty about their welfare, according to utilitarianism. But I should note two caveats. First, even if we might not be required to prioritize bringing about insects and other small animals in the near future, we might still be required to consider them much more than we do. For instance, to the degree that we expect practices such as agriculture, deforestation, and development to create more or fewer small animals with good or bad lives, we should assign weight to this consideration when deciding what to do.

Second, and additionally, even if humans are morally permitted to prioritize bringing about ourselves and other large animals in the near future, we might not be morally permitted to do so in the far future, according to utilitarianism. As I have mentioned and will discuss more in the next section, insofar as we take a longtermist perspective, it might be that we should prioritize bringing about very different kinds of beings in the far future. And either way, it might be that we should help insects and other small animals much more than we are at present as a means to this end, since helping these animals now is part of what will allow humanity to expand our moral circle and treat a wider range of nonhuman beings better in the future.

4. Pascal's Bugging

Utilitarianism might also have surprising implications about our duties regarding future microbe populations. In my view, the real test is not beings who have, say, a >10% chance of being sentient, but rather beings who have, say, a >.000001% chance of being sentient. Specifically, the real test is the prospect of an *astronomically high* number of beings who have an *astronomically low* but non-zero chance of being sentient. In this kind of case, as long as this population is large enough, utilitarianism (coupled with standard decision theory) can imply that we should favor it all else equal, even if its members all have a *very low* chance of being sentient and a *very low* level of welfare if any at all.

This problem is an instance of what Eliezer Yudkowsky calls *Pascal's mugging* (Yudkowsky, 2007 see also Bostrom, 2009). Suppose that a mugger walks up to you and proposes a deal: If you give them $10 today, then they promise to invest your money, make a profit, and give you $20 tomorrow. Plausibly, you have reason to decline this offer, since you have reason to doubt each part of this proposal: that the mugger would invest your money, that they would make a profit even if they did invest it, and that they would

pay you back even if they did make a profit. And as long as your credence that the mugger will keep this promise is lower than .5, giving them $10 is not worthwhile in expectation, all else equal.

So far so good. But now suppose the mugger keeps negotiating. They promise to return with increasingly high amounts of money, and they explain to you how they would convert your money into these higher amounts. With each new promise, you might think that the mugger is less likely to keep this promise but that it would be better if they did. In this case, as long as the increase in benefit sufficiently outpaces the decrease in probability (and the probability remains non-zero), at some point you might become rationally required to take the deal. This implication presents a challenge to standard decision theory. Can we really be rationally required to bet on an astronomically low chance of an astronomically high benefit?

Similar problems can arise for utilitarianism. Suppose we have an astronomically low chance of producing an astronomically high number of distant future happy people. In this case, as long as this population is large enough, we might be morally required to prioritize producing this population according to utilitarianism, all else equal. That is, utilitarianism might require us to pursue a course of action that has an astronomically low chance of creating and benefiting an astronomically high number of far future people, rather than courses of action that have a much higher chance of creating and/or benefiting much lower (but still high) numbers of current or near future people (for relevant discussion, see Greaves & MacAskill, 2021).

Now consider a similar case. Suppose that there is an astronomically low chance that microbes are sentient, and that each microbe can experience an astronomically low amount of welfare if they are. As before, as long as this population is large enough, we might be morally required to prioritize this population according to utilitarianism, all else equal. The math in this case is the same as the math in the previous case. The only difference is that, in the previous case, the subjects in question are definitely sentient but very unlikely to exist, whereas in this case, the subjects in question definitely exist but are very unlikely to be sentient (and have a very low capacity for welfare, if any at all). We can call this implication *Pascal's bugging*.

We can also imagine cases that involve both features at once. For instance, suppose that there is an astronomically low chance that we can bring about an astronomically high number of, say, *distant future microbes*. Suppose further that each microbe has an astronomically low chance of being sentient and has an astronomically low amount of welfare if any at all. Finally, suppose that we can still maximize expected utility by pursuing this option, given the sheer size of this population. In this case, a utilitarian might be required to forego creating and benefiting beings who are *very likely* sentient and/or who will *very likely* exist, all for the sake of beings who are merely *possibly* sentient and who will merely *possibly* exist.

As these examples illustrate, Pascal's bugging raises the stakes of the rebugnant conclusion in two related ways. First, it raises the stakes for intragenerational priority-setting questions because it implies that our decisions about how to set priorities within any particular generation depend not only on the implications for, say, insects, but also on the implications for even larger populations of even smaller beings. For example, if we discover that the world contains so many microbes that the 'micro-population' has more expected welfare than the 'macro-

population', then we would have to accept that the rightness or wrongness of our actions depends primarily on the expected impacts on the micro-population, all else equal.

Second, Pascal's bugging raises the stakes for intergenerational priority-setting because it implies that we might need to prioritize not only creating large populations of beings with a low amount of welfare on average, but also *very* large populations of beings with a *very* low chance of being sentient at all and a *very* low amount of welfare on average, if any at all. For example, if our choice is between creating a given number of future mammals, a quintillion times as many future insects, and a septillion times as many future microbes, then, depending on the probabilities and utilities we assign in each case, we might be morally required select not only the insects instead of the mammals but also the microbes instead of the insects.

As before, when faced with this kind of conclusion, we have several options. First, we can accept the conclusion. If current or future micro-populations did, in fact, contain more expected welfare than current or future macro-populations, then we would, in fact, be required to prioritize these mico-populations all else equal. Granted, this implication might seem highly implausible. But as noted above, every moral theory has some implications that we might find implausible, and the point of moral theory is to improve our moral thinking, not to merely confirm what we already think. And in this case, we can expect that human bias and ignorance would prime us to find Pascal's bugging highly implausible even if it was correct.

Alternatively, we can reject features of utilitarianism or standard decision theory that lead to this conclusion. For example, we can once again consider rejecting either the hedonism, totalism, or maximization of utilitarianism. Alternatively, we can consider rejecting the *fanaticism* of standard decision theory, according to which we should consider all possibilities when deciding what to do, including possibilities involving astronomically low probabilities of astronomically high impacts. Rejecting this idea might allow us to discount or bracket astronomically low probabilities of astronomically high impacts when estimating how to do the most good possible (for discussion, see Wilkinson, 2022).

Of course, as with the rebugnant conclusion, even if we accept Pascal's bugging in principle, we might still be able to reject or constrain it in many cases in practice. For example, if we think that we have no knowledge or power regarding microbe welfare, then we might think that we can bracket our impacts on microbes in practice *whether or not* they have more expected value than other animals. But a lot depends on whether we have *zero* knowledge and power or, rather, *very little* knowledge and power. If we have *zero* knowledge and power, then this reply might work. But if we instead have *very little* knowledge and power, then while this reply might add an extra line to our math, it might or might not alter the outcome.

Similarly, if we think that the probability that microbes are sentient is *zero*, then we might be able to avoid this problem with respect to microbe populations. Or if we think that the probabilities and utilities are so low that other populations still have more expected welfare in total, despite the size of microbe populations, then we might be able to avoid the problem with for this reason as well. Given the problem of other minds, I think that the 'low expected welfare' response is more plausible than the 'no expected welfare' response. It would be rash to think that there is *no chance at all* that microbes can

consciously experience *any positive or negative states at all*, no matter how minimal. But both responses can be explored.

While there is too much uncertainty for us to say much with confidence at this stage, we have at least some reason to think that microbe populations could carry a lot of weight in expectation. For instance, according to one recent estimate, 'there are between 10^23 and 10^24 neurons on earth', most of which are 'distributed roughly evenly among small land arthropods, fish, and nematodes, or possibly dominated by nematodes'.[10] Nematodes are microscopic organisms with about 300 neurons each. In my view, they are good candidates for the kind of being that we are considering: very low but non-zero chance of being sentient, very low capacity for welfare if any at all, but numerous enough that it could all add up.

As before, we might or might not think that we should *prioritize* nematodes in practice any time soon, if any time at all, according to utilitarianism. But even if not, we might still think that we should at least *consider* them (along with many other possibly sentient beings, as we will see in a moment) much more than we are. In this case we can make all the same points as before: A lot depends on the tractability and indirect effects of promoting nematode welfare, as well as on the exact probabilities and utilities involved. But if and when we do determine that nematodes have net positive or negative expected welfare, we might once again have to treat any policy that creates more or fewer of them as good or bad to that degree.

5. Bugs in the System

Throughout this discussion I have emphasized that utilitarianism might have these implications not only in principle but also, at least to a degree, in some cases in practice. I now want to consider two possible scenarios where that might be true. In the first scenario, utilitarianism could imply that the rightness or wrongness of our actions depends primarily on the expected impacts for large populations of small animals. In the second scenario, it could imply that the rightness or wrongness of our actions depends primarily on the expected impacts for large populations of AI systems. Both of these scenarios are possible, and they could at least partly match the scenarios that we have been considering.

Consider first the scenario involving small animals. Many humans think that global changes such as climate change are bad because of the expected impacts for humans. Of course, this view makes sense. Climate change will cause melting ice caps, rising sea levels, flooding coastal areas, conflict over land, water, and energy, and an increase in extreme weather events such as hurricanes and tsunamis. It will also be a threat multiplier that amplifies existing threats for humans such as hunger, thirst, illness, and injury. As a result, many humans will suffer and die, including and especially the most vulnerable among us. While many humans will also prosper, we can reasonably expect that the net effects for humans will be negative.[11]

But when we consider nonhumans too, we realize that the impacts of climate change will be much more complex. While many animals will suffer, many other animals will flourish. And while many nonhuman populations will contract, many others will expand. While much remains unclear, one possibility – which, to be clear, is only a possibility – is that climate change will produce a world with a higher ratio of small to large animals.

Warmer climates tend to favor smaller animals, both by allowing them to migrate north and south and by causing some larger animals to shrink. As a result, a world reshaped by climate change could contain more animals and higher welfare in total but smaller animals and lower welfare on average.[12]

If this scenario arises – which, again, is a big 'if' – then a lot will depend on whether these small animals have net positive or negative expected welfare. If they have net positive expected welfare, then climate change could produce a real-life rebugnant conclusion, by producing larger populations of smaller animals, such that they experience less pleasure on average but more in total. Otherwise it could have the reverse effect. Either way, if these small animals have more expected welfare than humans and other large animals during this period, then the goodness or badness of climate change during this period will depend primarily on the impacts for these small animals, according to utilitarianism.[13]

Now consider the scenario involving AI systems. If humanity survives climate change and the other global threats that will imperil our species in the near term, then we might face an open future with opportunities for expansion. Specifically, we might have the ability to expand beyond the planet, and we might also have the motivation to do so. After all, as long as we can survive as a species, we likely have at least a billion years before water loss and thermal runaway make this planet uninhabitable (Wolf & Toon, 2014). This gives us plenty of time to develop the knowledge and power necessary to expand. We then have billions of years to spend potentially extending sentient life throughout the universe (Bostrom, 2003).

The question is how we would go about extending sentient life throughout the universe, if we could. Consider two options. The first is a biological approach: We aim to create as many happy biological minds as possible, by building space stations or terraformed planets on which (post-)human and nonhuman animals can live. The second is an artificial approach: We aim to create as many happy artificial minds as possible, by building large, solar-powered networks of artificial minds who can experience pleasure. While we can plausibly achieve a very large population either way, we can plausibly achieve a much larger population and much better outcomes for its members with the artificial approach (John, 2021).

I imagine that utilitarian agents might pursue a mixed approach at first, if for no other reason than to further reduce existential risk and oversee whatever further expansions might be optimal. But I can also imagine scenarios where utilitarian agents pursue an artificial-centric approach beyond that. This approach might or might not fit the structure of the rebugnant conclusion in at least some respects. For instance, depending on the details, it might or might not be that future artificial minds have a lower chance of being sentient and a lower capacity for welfare than future biological minds. But even if they do, the astronomically large size of a future artificial population could be more than enough to make up for that.

Somewhat ironically, it might be that utilitarians can avoid the first 'repugnant' conclusion in part by embracing the second one. That is, even *if* we were to determine that climate change is net positive in the short term (because it creates more small animals and higher total welfare), we might still determine that climate change is net negative in the long run (because it increases existential risks for humanity, thereby reducing the probability that we can survive long enough to spread positive welfare beyond this planet). In

this scenario, utilitarianism would vindicate the commonsense view that climate change is bad, but it might do so only instrumentally, in service of strange and surprising far future goals.

With that said, as I have emphasized throughout this paper, even if utilitarians should accept these 'repugnant' conclusions in theory, we might be able to reject some of them in practice. As noted above, utilitarians should accept a pluralistic moral framework that considers welfare, rights, virtues, relationships, and more in practice. Moreover, we need to consider many factors in our impact assessments in practice, such as: how likely particular beings are to be sentient, how much welfare these beings are likely to have, to what extent particular actions will increase or decrease their welfare, and to what extent our successors can overcome epistemic, practical, and motivational obstacles that stand in the way of impartiality.

Once we consider all these factors, we might decide that utilitarians are unlikely to face pure versions of the rebugnant conclusion, Pascal's bugging, or other such problems in practice. For instance, we might decide that promoting welfare while respecting rights, cultivating virtuous characters, and cultivating caring relationships requires not making high-stakes decisions through classical utilitarian reasoning alone. And even insofar as we apply classical utilitarian reasoning in these situations, we might decide that climate change, the artificial path, and other such outcomes are not, in fact, better in total but worse on average, in expectation. In these scenarios, we might still face a strange moral future, but the details might seem more plausible.

But we should keep in mind that, insofar as we do avoid a 'repugnant' conclusion, we might face another surprising conclusion instead. In particular, when we ask what kind of future population is best all things considered, the answer will likely *not* be an expanded population of beings like us, but will likely instead be *either* (a) a much larger population of much smaller beings *or* (b) a much smaller population of much larger beings. And whereas answer (a) might remind us of a 'repugnant conclusion' case, answer (b) might remind us of a 'utility monster' case, that is, a case where we can maximize expected utility by prioritizing a relatively small number of beings with a relatively large amount of happiness on average (Nozick, 1974).

We should also keep in mind that, even if an expanded population of beings like us *is* best all things considered, that fact would still seem like a 'repugnant conclusion' case to a much smaller population of much larger beings, and it would still seem like a 'utility monster' case to a much larger population of much smaller beings. While we might be inclined to see human-like populations as the norm and other kinds of populations as deviations, the reality is that every kind of population is a norm from some possible perspectives and a deviation from others. And if we want to think about population ethics in an impartial manner, then we should avoid anchoring to our own contingent perspectives about such matters.

Granted, even if we keep all these considerations in mind, we might still have a hard time shaking the intuition that we should prioritize human-like populations in both the short term and the long run. No matter how much progress we make as a species, we might continue to find it strange that our reward for improving human lives and social, political, and economic systems is that we have an increasingly strong responsibility to live in service of very different kinds of (possibly) sentient being. But these strike me as the kinds of bullets that utilitarians should be willing to bite, especially since we know that

human bias and ignorance will likely continue to shape our moral intuitions even as our beliefs, values, and practices improve.

In any case, I think that the immediate implications of this discussion are highly intuitively plausible: We should work at present to build knowledge, power, and political will toward creating a better world for humans and other sentient beings. We might or might not face any pure 'repugnant' or 'utility monster' cases in practice. But if we do, we will need to be epistemically, practically, and motivationally prepared. And even if not, we will still face the weaker – but still radical – conclusion that beings who are very different from us might matter much more than we appreciate. We will need to consider our impacts on these beings very carefully in the future, whether or not we should prioritize them over ourselves.

In closing, it is worth reiterating that many of these problems can arise for non-utilitarian moral theories as well. Whether we think about morality primarily in terms of welfare, rights, virtues, relationships, or other features of life, we will have to ask how to set priorities between large populations of small beings and small populations of large beings, both at present and in the future. And while we might be able to avoid some of the implications discussed here by rejecting features of utilitarianism such as hedonism, totalism, maximization, or fanaticism, we might not be able to avoid all of them that way. Ultimately, the more we accept how large and varied the moral community is, the stranger morality will become.

Notes

1. Thanks to Kyle Johannsen, two anonymous referees, and the organizers and participants of the 8th Oxford Workshop on Global Priorities Research for helpful feedback on previous drafts of this paper.
2. Note that some people use the term 'welfare' in a valenced way, to refer to positive welfare states (in this case we can contrast 'welfare' with 'illfare'), whereas other people use the term 'welfare' in a non-valenced way, to refer to any welfare states (in this case we can contrast 'positive welfare' with 'negative welfare'). This paper will use the term in this latter sense, but nothing substantive will turn on that terminological choice.
3. For more on the features of utilitarianism discussed in this section, see De Lazari-Radek and Singer (1974). For general discussion of the pros and cons of utilitarianism and other moral theories, see Driver (2007).
4. This paper will use 'utility' to refer to goodness in this hedonic sense.
5. For more on this kind of indirect consequentialism, see Sidgwick (1874/2011 and Hare (1981).
6. For similar discussion, see Sebo (2022).
7. For analysis of key assumptions underlying totalism, see Hirose (2014).
8. For discussion of non-maximizing views, see Jamieson and Elliot (2009), Slote (1984), and Sinhababu (2018).
9. For discussion of why the repugnant conclusion is not necessarily repugnant, see Zuber et al. (2021).
10. For informal discussion, see Ray (2017), 'How many neurons are there?' *Eukaryote Writes Blog*: https://eukaryotewritesblog.com/how-many-neurons-are-there/.
11. For more on the nature and ethics of climate change, see Jamieson (2014).
12. For more on the impacts on the distribution of animals, see Bebber et al. (2013). For more on the impacts on the size of animals, see Weeks et al. (2020).
13. We might also think that large populations of small animals are more likely than small populations of large animals to have net negative welfare, since the former populations are more likely to be r-selected and, so, to have higher infant mortality rates. For more, see Johannsen (2021).

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Bebber, D., Ramotowski, M., & Gurr, S. (2013). Crop pests and pathogens move polewards in a warming world. *Nature Climate Change*, *3*(11), 985–988. https://doi.org/10.1038/nclimate1990

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*(3), 308–314. https://doi.org/10.1017/S0953820800004076

Bostrom, N. (2009). Pascal's wager. *Analysis*, *69*(3), 443–445. https://doi.org/10.1093/analys/anp062

Carruthers, P. (2004). *The Nature of the mind: An introduction*, *Chapter 1*. Routledge.

De Lazari-Radek, K., & Singer, P. (2017). *Utilitarianism: A very short introduction*. Oxford University Press).

Driver, J. (2007). *Ethics the fundamentals*. Blackwell Publishing.

Greaves, H., & MacAskill, W. (2021). *The case for strong longtermism* (GPI Working Paper No. 5). https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/

Groff, Z., & Ng, Y.K. (2019). Does suffering dominate enjoyment in the Animal Kingdom? An update to welfare biology. *Biology & Philosophy*, *34*(4), 40. https://doi.org/10.1007/s10539-019-9692-0

Hare, R. M. (1981). *Moral thinking: Its levels, method, and point*. Oxford University Press.

Hirose, I. (2014). *Moral aggregation*. Oxford University Press.

Horta, O. (2010). Debunking the idyllic view of natural processes: Population dynamics and suffering in the wild. *Telos: Revista Iberoamericana de Estudios Utilitaristas*, *17*(1), 73–90.

Jamieson, D. (2014). *Reason in a dark time: Why the struggle against climate change failed - and what it means for our future*. Oxford University Press.

Jamieson, D., & Elliot, R. (2009). Progressive consequentialism. *Philosophical Perspectives*, *23*(1), 241–251. https://doi.org/10.1111/j.1520-8583.2009.00169.x

Johannsen, K. (2021). *Wild animal ethics: The moral and political problem of wild animal suffering*. Routledge.

John, T. (2021). Panspecies longtermism, unpublished manuscript.

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Ng, Y.K. (1989). What should we do about future generations? *Economics & Philosophy*, *5*(2), 245–253. https://doi.org/10.1017/S0266267100002406

Ng, Y.K. (1995). Towards welfare biology: Evolutionary economics of animal consciousness and suffering. *Biology & Philosophy*, *10*(3), 255–285. https://doi.org/10.1007/BF00852469

Nozick, R. (1974). *Anarchy, state, and Utopia*. Basic Books.

Palmer, C. (2010). *Animal ethics in context*. Columbia University Press.

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Ray, G. (2017). *How many neurons are there?* Eukaryote Writes Blog. https://eukaryotewritesblog.com/how-many-neurons-are-there/

Sanders, B. (2020). *Global animal slaughter statistics and charts: 2020 update*. Faunalytics. https://faunalytics.org/global-animal-slaughter-statistics-and-charts-2020-update/

Schukraft, J. (2020). *Differences in the intensity of valenced experience across species*. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/H7KMqMtqNifGYMDft/differences-in-the-intensity-of-valenced-experience-across

Sebo, J. (2018). The moral problem of other minds. *The Harvard Review of Philosophy*, *25*, 51–70. https://doi.org/10.5840/harvardreview20185913

Sebo, J. (2022). *Saving animals, saving ourselves*. Oxford University Press.

Sidgwick, H. (1874/2011). *The methods of ethics*. Cambridge University Press.

Sinhababu, N. (2018). Scalar consequentialism the right way. *Philosophical Studies*, *175*(12), 3131–3144. https://doi.org/10.1007/s11098-017-0998-y

Slote, M. (1984). Satisficing consequentialism. *Proceedings of the Aristotelian Society*, *58*(1), 139–176. https://doi.org/10.1093/aristoteliansupp/58.1.139

Tomasik, B. (2015). The importance of wild animal suffering. *Relations*, *3*(2), 133–152. https://doi.org/10.7358/rela-2015-002-toma

Weeks, B. C., Willard, D. E., Zimova, M., Ellis, A. A., Witynski, M. L., Hennen, M., & Winger, B. M. (2020). Shared morphological consequences of global warming in North American migratory birds. *Ecology Letters*, *23*(2), 316–325. https://doi.org/10.1111/ele.13434

Wilkinson, H. (2022). In defence of fanaticism. *Ethics*, *132(*2): 445–477.

Wolf, E. T., & Toon, O. B. (2014). Delayed onset of runaway and moist greenhouse climates for earth. *Geophysical Research Letters*, *41*(1), 167–172. https://doi.org/10.1002/2013GL058376

Yudkowsky, E. (2007). *Pascal's mugging: Tiny probabilities of vast utilities*. Less Wrong. https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities.

Zuber, S., Venkatesh, N., Tännsjö, T., Tarsney, C., Stefánsson, H. O., Steele, K., Spears, D., Sebo, J., Pivato, M., Ord, T., Ng, Y.K., Masny, M., MacAskill, W., Lawson, N., Kuruc, K., Hutchinson, M., Gustafsson, J. E., Greaves, H., Forsberg, L., & Asheim, G. B. (2021). What should we agree on about the repugnant conclusion? *Utilitas*, *33*(4), 1–5. https://doi.org/10.1017/S095382082100011X