Inquiry

An Interdisciplinary Journal of Philosophy

ISSN: 0020-174X (Print) 1502-3923 (Online) Journal homepage: https://www.tandfonline.com/loi/sinq20

The world destruction argument

Simon Knutsson

To cite this article: Simon Knutsson (2019): The world destruction argument, Inquiry, DOI: 10.1080/0020174X.2019.1658631

To link to this article: https://doi.org/10.1080/0020174X.2019.1658631

Published online: 29 Aug 2019.

INQUIRY
https://doi.org/10.1080/0020174X.2019.1658631

Routledge
Taylor & Francis Group

The world destruction argument

Simon Knutsson

Department of Philosophy, Stockholm University, Stockholm, Sweden

ABSTRACT
The most common argument against negative utilitarianism is the world destruction argument, according to which negative utilitarianism implies that if someone could kill everyone or destroy the world, it would be her duty to do so. Those making the argument often endorse some other form of consequentialism, usually traditional utilitarianism. It has been assumed that negative utilitarianism is less plausible than such other theories partly because of the world destruction argument. So, it is thought, someone who finds theories in the spirit of utilitarianism attractive should not go for negative utilitarianism, but should instead pick traditional utilitarianism or some other similar theory such as prioritarianism. I argue that this is a mistake. The world destruction argument is not a reason to reject negative utilitarianism in favour of these other forms of consequentialism, because there are similar arguments against such theories that are at least as persuasive as the world destruction argument is against negative utilitarianism.

ARTICLE HISTORY Received 16 January 2019; Accepted 18 July 2019

KEYWORDS Negative utilitarianism; consequentialism; utilitarianism; killing; replacement

1. Introduction

Negative utilitarianism is often understood as the moral theory whose only prescription is that we should minimise suffering or negative well-being, and that is the conception I will assume here.[1] The most discussed argument against negative utilitarianism is roughly this: Negative utilitarianism implies that one should kill all humans or all sentient life, or destroy the world, if one had the opportunity. Doing so would be wrong, and hence the plausibility of negative utilitarianism is undermined. I call it 'the world destruction argument', but will for brevity's sake mostly refer to it as 'the elimination argument'.[2]

CONTACT Simon Knutsson simon.knutsson@philosophy.su.se; simonknutsson@gmail.com

[1] This is a *strong* form of negative utilitarianism. *Weak* versions give weight to both positive and negative well-being, but more weight to negative well-being (Griffin 1979; Arrhenius and Bykvist 1995).

[2] The phrase 'the elimination argument' is from Arrhenius and Bykvist (1995, 31).

In 1955, Ingemar Hedenius made this argument in Swedish against his own form of consequentialism, according to which goods cannot counterbalance some evils.[3] An English formulation followed in 1958 by R. N. Smart who argued against negative utilitarianism.[4] The argument is often mentioned in applied and interdisciplinary writings,[5] and it has been endorsed by philosophers such as J. J. C. Smart, Rem B. Edwards, Mario Bunge, David Heyd, Gustaf Arrhenius, Krister Bykvist, and, as recently as 2013 and 2015, Toby Ord and Torbjörn Tännsjö.[6]

The elimination argument against negative utilitarianism is different from the general objection that negative utilitarianism is implausible because it gives too much weight to suffering. An important feature of the elimination argument is that it concerns violence and the acts of killing and destroying. For example, Ord writes that 'a thorough going Negative Utilitarian would support the destruction of the world (even by violent means)'.[7] Such phrases have played an important role in the case made against negative utilitarianism during the last six or so decades. The phrases essentially paint a picture of the negative utilitarian as a dangerous fanatic. The objection is not merely that negative utilitarianism includes an implausible value theory or implies that it would be right to peacefully give up arbitrarily large amounts of positive well-being to avoid much smaller (even trivial) amounts of negative well-being. Therefore, I will focus on the act of killing everyone rather than on whether negative utilitarianism, in general, gives too much weight to suffering. That said, there are connections that I will touch upon between how objectionable it would be to kill everyone and how plausible a theory's weighing of positive versus negative well-being is.

Those making the elimination argument against negative utilitarianism often express sympathy for some other form of consequentialism that does not emphasise the reduction of suffering as much as negative utilitarianism does. Usually, sympathy is expressed for traditional utilitarianism – that is, a form of utilitarianism in which positive and negative well-being have equal weight or importance. It has been assumed that negative

[3] Hedenius (1955, 45, 100–105).

[4] R. N. Smart (1958).

[5] E.g. Robertson, Morris, and Walter (2007, 404), and Baum and Wilson (2013).

[6] J. J. C. Smart (1973, 29), Edwards (1986, 133), Bunge (1989, 230), Heyd (1992, 60), Arrhenius and Bykvist (1995, sec. 4.2), Ord (2013), and Tännsjö (2015, 243–44). Strictly speaking, Heyd does not mention killing others as a path to 'the painless annihilation of all humanity'; only collective suicide and abstention from procreation. Tännsjö appears to endorse R. N. Smart's elimination argument against negative utilitarianism because he mentions it and shortly thereafter concludes that one should not give suffering lexical weight.

[7] Ord (2013, sec. History).

utilitarianism is less plausible than such other theories partly because of the elimination argument. I argue that this is not so. The elimination argument is not a basis for rejecting negative utilitarianism in favour of these other forms of consequentialism, because there are similar arguments against such theories that are at least as convincing as the elimination argument is against negative utilitarianism. For example, Dale Jamieson wrote as early as 1984 that

> many philosophers have rejected TU [total utilitarianism] because it seems vulnerable to the Replacement Argument and the Repugnant Conclusion. … The Replacement Argument purports to show that a utilitarian cannot object to painlessly killing everyone now alive, so long as they are replaced with equally happy people who would not otherwise have lived. (Jamieson 1984, 218)

This replacement argument against traditional utilitarianism has barely been mentioned in the academic literature since 1984, while the similar elimination argument against negative utilitarianism continues to be cited.

I aim to compare the plausibility of negative utilitarianism with other forms of consequentialism, such as traditional utilitarianism and prioritarianism (and even with some theories that are not purely consequentialist, although that is peripheral). More specifically, I am interested in whether negative utilitarianism is more vulnerable than these other theories are to 'elimination arguments', by which I mean arguments that are similar to the elimination argument presented above. This is important for those who are attracted to theories in this cluster and who are deciding which theory to choose. For brevity and simplicity, I only investigate which form of utilitarianism – negative or traditional – is more vulnerable to elimination arguments, and I conclude that elimination arguments are at least as persuasive against traditional utilitarianism as against negative utilitarianism. But my investigation also sheds light on how vulnerable negative utilitarianism is compared to other theories besides traditional utilitarianism. My investigation generalises more easily to other theories the more similar they are to traditional or negative utilitarianism in the relevant respects. For example, it generalises well to prioritarianism. Indeed, one could argue along the lines of this paper that negative utilitarianism is no more vulnerable to elimination arguments than some theories that are not even purely consequentialist, such as a pluralist theory that includes both consequentialist and virtue ethical ideas. Such a theory could hold that one ought to maximise final value; well-being has final value and is aggregated as in traditional utilitarianism; beauty also has final value; and courage is a virtue that one should have, even if being courageous does not lead to better consequences. This pluralist theory could be accused of entailing that if one could, in a way consistent with a courageous character, eliminate the world and replace it with a more beautiful world with more positive well-being, one should do so.

My overall conclusion is that those who claim that negative utilitarianism is inferior to such other views because of elimination arguments need to use another argument or show, in more detail than they have done, that their preferred theory handles elimination arguments better than does negative utilitarianism.

I will focus on negative versus traditional *total act*-utilitarianism for simplicity, and because previous discussions of the elimination argument have mainly contrasted these two theories. I will henceforth understand negative utilitarianism and traditional utilitarianism as the following criteria of rightness:

> *Negative total act-utilitarianism*: An act is right if and only if it results in a sum of negative well-being that is at least as small as that resulting from any other available act.

> *Traditional total act-utilitarianism*: An act is right if and only if it results in a sum of well-being (positive well-being minus negative well-being) that is at least as great as that resulting from any other available act.

Since it matters little to my main points whether one formulates these theories in terms of actual or expected results, I will sometimes speak of expected results or expected value, and sometimes simply of results. For brevity, I will often speak of suffering instead of negative well-being. I will concentrate on individual rather than group agents because for some group agents the investigation would look quite different (for example, if a state or a world government is taken as a single agent).

2. Elimination arguments and types of replies to them

Others have brought up at least five types of elimination arguments in the context of negative or traditional utilitarianism. I here rephrase them somewhat and give them new names, except the name ‘Elimination’, which is the term used in Arrhenius and Bykvist (1995, 31). The following are the two arguments against negative utilitarianism:

> *Elimination*: Someone can kill all humans or all sentient beings on Earth painlessly. Negative utilitarianism implies that it would be right to do so.[8]

[8] See Hedenius (1955, 45, 100–105) and R. N. Smart (1958).

> *Paradise with Suffering*: The world has become a paradise, yet would contain some (possibly mild and brief) suffering if it remained. Someone can kill everyone in this paradise instantly and painlessly. Negative utilitarianism implies that it would be right to do so.[9]

The following are the three arguments that pertain to traditional utilitarianism:

> *Traditional Utilitarian Elimination*: The sum of positive and negative well-being in the future will be negative if humans or sentient life continues to exist. Traditional utilitarianism implies that it would be right to kill all humans or all sentient beings on Earth painlessly.[10]

> *Suboptimal Earth*: Someone can kill all humans or all sentient beings on Earth and replace us with new sentient beings such as genetically modified biological beings, brains in vats, or sentient machines. The new beings could come into existence on Earth or elsewhere. The future sum of well-being would thereby become (possibly only slightly) greater. Traditional utilitarianism implies that it would be right to kill and replace everyone.[11]

> *Suboptimal Paradise*: The world has become a paradise with no suffering. Someone can kill everyone in this paradise and replace them with beings with (possibly only slightly) more well-being in total. Traditional utilitarianism implies that it would be right to do so.[12]

I have formulated the arguments in terms of killing rather than world destruction because it is simpler and more realistic, and because this difference matters little when one compares negative and traditional utilitarianism since according to both theories, only well-being has final value.

Elimination arguments have been discussed little compared to other, well-known, smaller-scale counterexamples to utilitarianism. Such examples include the doctor who can kill one patient to harvest her organs and give them to five others; the sheriff who can frame and execute an innocent person to prevent riots; and replacement cases such as killing animals in the animal industry or some infants who are replaced with new ones with at least as much well-being (McCloskey 1957, 468–69; Thomson 1976, 206; Pluhar 1990; Mulgan 2007, 94–95).

[9] See Pearce (2005).

[10] J. J. C. Smart writes, 'A classical utilitarian could be a benevolent world exploder only if he or she were a pessimist who, like Schopenhauer, believed that sentient beings inevitably, or perhaps even for the most part, are more miserable than happy' (1989, 43). This statement is problematic, as for classical and traditional utilitarianism to imply that killing everyone would be right, currently existing beings need not be more miserable than happy, and nothing close to Schopenhauer's pessimism is required.

[11] See Jamieson (1984, 218) and Pearce (2013).

[12] See Pearce (2013).

Utilitarians have given at least three kinds of replies to these smaller-scale counterexamples, which traditional and negative utilitarians could attempt to use as replies to elimination arguments. I will investigate whether these replies, when used as replies to elimination arguments, are more convincing when offered in defence of traditional utilitarianism versus in defence of negative utilitarianism. One such reply is that killing would not be so objectionable or counterintuitive, or at least less objectionable than the other available actions or the implications of competing moral theories (J. J. C. Smart 1973, 71–73). Another reply is that killing is unlikely to be the optimal act in real life (Sprigge 1965, 275–78; Hare 1981, 134, 163–4). The final reply is to appeal to indirect act-utilitarianism (Hare 1981, 132–35).

3. How objectionable are the purported implications?

Elimination arguments are meant to show that a moral theory has objectionable or counterintuitive implications. Those who use elimination arguments against negative utilitarianism might argue that their theories have the upper hand because, even though they also imply that it would be right to destroy the world or kill everyone in some situations, there are fewer such situations or killing everyone in those situations is less objectionable than doing so in the situations in which negative utilitarianism recommends killing everyone. In this section, I will consider some ways to support and counter such arguments. In the end, I find the purported implications of traditional utilitarianism more objectionable than those of negative utilitarianism.

First, in defence of traditional utilitarianism, one can say that at least the theory only implies that it would be right to kill everyone in a proper subset of all scenarios in which someone could kill everyone. It would only be right if it would result in a sum of well-being that is at least as great as that resulting from any other available act. A reply in defence of negative utilitarianism is that it too merely implies that killing everyone would be right in a proper subset of scenarios in which someone could kill everyone. There are scenarios in which traditional utilitarianism, but not negative utilitarianism, implies that it would be right to kill everyone, namely, scenarios in which the killing would increase both positive and negative well-being and result in a greater sum of positive minus negative well-being. Negative utilitarianism does not imply that it would be right to kill everyone in such scenarios because, in these scenarios, killing everyone would increase negative well-being. An example of such a scenario is that all humans or all sentient beings on Earth could be killed and replaced with many more beings who, collectively, experience both more positive well-being and more negative well-being, but with a greater sum of positive minus negative well-being. According to this reply, traditional utilitarianism does imply that in this scenario it would be right to kill everyone currently alive. (I describe such scenarios in more detail in the latter part of section 4 below.)

Second, in favour of traditional utilitarianism, one may argue that at least it implies that killing everyone would be right in fewer scenarios or a smaller share of possible scenarios than negative utilitarianism does. However, such talk of 'fewer' or 'smaller share' is too unclear. The argument would need to be supplemented with a specification of how one counts the number or share of possible scenarios.

Third, the negative utilitarian can argue that losing what currently exists on Earth would not be much of a loss, because of the following very plausible observation: overall, sentient beings on Earth fare terribly badly. The situation is not terrible for every single individual, but it is terrible when all individuals are considered. We can divide most sentient beings on Earth into the three categories of humans, non-human animals in captivity, and wild non-human animals, and deal with them in turn. Benatar (2017, 91) argues that 'the quality of human life is not only much worse than most people think but actually quite awful'. He points out how much of human life is permeated by unpleasant experiences, frustrated desires and so on, and he argues that most people's self-assessments of their quality of life are unreliable due to biases (chap. 4). I would also emphasise the worst cruelty that some humans suffer (see, e.g., Matthews 2008; Mukwege and Nangini 2009). As for animals in captivity, it is easier to argue that they generally lead bad lives, as most of them are held and killed in the animal industry for the production of food and other products. Agricultural economist F. Bailey Norwood and Sara Shields, an animal welfare expert, estimate the lifetime welfare of US farm animals on a scale from −10 to 10. I list their scores as pairs, where the first score in the pair is Norwood's and the second is Shields'.[13] Their ratings are as follows: cows raised for meat 6, 2; dairy cows 4, 0; chickens raised for meat 3, −8; pigs −2, −5; and egg-laying hens in cage systems −8, −7. Shields rates the welfare of fish −7. Importantly and unfortunately, the most numerous of these animals (the chickens, hens and fish) tend, with some exceptions, to get

[13] For brevity, I omit some of Norwood's estimates, such as his score of 4 for 'shelter-pasture pork'. All scores by Norwood that I list are for market (non-breeder) animals, which I presume is also the case for Shields' scores.

scores around −7 and −8.[14] These are merely judgment calls by two experts, and especially Norwood's scores seem too optimistic. My view after having studied the animal industry is that, in general, the animals' situation is appalling. Finally, consider the well-being of wild animals; this also seems to be generally poor. One important argument for this conclusion in the literature is that it is common for animals to have many offspring, most of whom die young (Ng 1995, 270–72; Horta 2010). Add to this the extent of starvation, violence, disease and injury among wild animals and the picture becomes even grimmer.[15]

Fourth, one could try to convince the reader that it is intuitive or not so objectionable that it would be right to kill everyone in some situations by informing the reader of what the long-term future may be like or by painting a vision. Traditional utilitarian Tännsjö has written favourably about *replacing* all humans, but not, to my knowledge, about *killing* and replacing all humans. He may, for example, have a peaceful, gradual replacement in mind. In any case, one could use what he says to defend killing and replacing everyone. He writes (my translation),

> Our individual death is instrumentally valuable. It is good that we step down in favour of new generations, who can see the world with new, fresh eyes. … After me come other happy beings, who better than I, when I have become old and tired, can enjoy their lives. The replacement is a condition for a dynamic continuation of humanity's history. … Suppose that we really can replace humans with beings who are wiser, more able to live in peace and harmony with one another, and far more innovative and with a great appetite for life. Suppose that they can live far better lives than the lives we live; suppose that they can handle different negative existential threats far better than we can. Why would "we" then not step down in favour of them? I think that it is clear that this is what in this situation *should* happen. … I guess that it starts with genetically improved beings. … The development will continue with artificial intelligences. … By their construction … there will be room made for an exponential growth of the number of happy beings in the universe, who live in mutual symbiosis without getting in one another's way. Imagine a sparkling starry sky, where each light that twinkles towards you is a blissful robot. And behind each light, new skies. … Let us rejoice with all those who one day hopefully … will take our place in the universe.[16]

What he writes might make some people feel better about being replaced and perhaps even think that killing everyone to replace us would not be so

[14] Norwood and Lusk (2011, 229) and Cooney (2013, 7). An exception is that Norwood gives laying hens in cage-free systems the score 2.

[15] The organisation Animal Ethics has published texts such as 'Malnutrition, hunger and thirst in wild animals' and 'Diseases in nature', which are available at Animal Ethics (2016).

[16] Tännsjö (2016). He speaks similarly in Tännsjö (1997, 245).

terrible, but a way to make progress. In response, a negative utilitarian could also paint a picture and provide facts in support of how bad the future may be. Consider all the gruesome violence and suffering that will almost certainly persist in the near term if we survive. Even worse, imagine the horror that may be realised in the farther future. For example, according to Peters (1985, 172), torture methods of the late twentieth century 'produce a range and intensity of pain that greatly exceeds that of earlier forms of torture'. It is easy to imagine that future torture will be much worse. Why should we not step down, stop the perpetuation of extreme suffering on Earth and spare future victims from coming into existence?

Fifth, perhaps the strongest argument a negative utilitarian can present in favour of the claim that killing everyone would not be so objectionable is that killing to reduce suffering is regrettable but still the lesser of evils in the sense of resulting in less disvalue than other options. This argument is not available to traditional utilitarians when the killing of everyone would result in more negative and positive well-being and a greater sum of positive minus negative well-being. A traditional utilitarian may reply that in this case killing would be the lesser evil because not bringing about the positive well-being would be the greater evil in the different sense of resulting in less positive minus negative well-being than other options. My reply is that killing to prevent more bad things from happening is one of the types of killing that seem least objectionable in general (another type is killing in self-defence). I have in mind cases such as euthanising injured or sick animals to end their suffering, a police officer shooting someone who is about to seriously harm or kill members of the public, or killing in war to prevent more violence. In contrast, to bring about new beings with a greater total of positive well-being does not seem sufficient to warrant such a serious act as killing.

Finally, in favour of negative utilitarianism, one can argue that killing everyone would not be so objectionable because what ultimately matters is only the reduction of negative well-being. Analogously, one can defend traditional utilitarianism by saying that it would not be so objectionable to kill everyone and replace us with other beings who will experience a greater sum of positive minus negative well-being because our only obligation is to maximise the sum of positive minus negative well-being. Which of these positions is more plausible is a larger discussion mainly beyond the scope of this paper. Here, I will merely reply to some common critical questions about the extent to which negative utilitarianism focuses on the negative (a negative utilitarian can, but need not, reply as I do). Hopefully, this exercise will show why someone might, as I do, find it appealing to give no moral weight to positive well-being. This may, in turn, affect how objectionable the purported implications of traditional and negative utilitarianism are. Let us start with the first critical question. Why do you care only about negative well-being and not at all about positive well-being? Because there is no positive well-being, and there will never be any.[17] Common notions of an individual's positive well-being include that it is or concerns what is good for the individual; that is, what has positive final value for the individual. Here, I make the axiological claim that nothing has positive final value for individuals. We can continue saying in daily life that it was good for someone that they became healthier or that they had some experience, but when saying so, I would not mean that these things had positive final value for the person; I would mean that the person would otherwise have been worse off. Despite my claim that positive well-being does not exist, I maintain that negative well-being exists, which is also an axiological claim. It is, for example, bad for someone to have the experience of being tortured.

Let us consider further critical follow-up questions. I find it plausible that negative experiences have negative value for an individual; more specifically, it is plausible that some experiences have a negative hedonic tone (quality) and that they are bad for the individual who has the experiences. A related form of hedonism is that experiences have a negative, neutral, or positive hedonic tone. Now, someone might object, if there are experiences with a negative hedonic tone that are bad for an individual, why are experiences with a positive hedonic tone not good for an individual? Because there are no experiences with a positive hedonic tone, and there will never be any. To be clear, I do not deny the existence of pleasure in the Epicurean sense of katastematic (static) pleasure, which includes tranquillity and the absence of pain, trouble and worry, and which 'can be varied, though not increased' (Annas 1993, 188, 336). In daily life, it is common to use phrases such as 'this is very pleasant', which is fine; I do not object to that usage of words. If I were at a spa, I would perhaps say that my experience is pleasant, but I would not mean that the experience has a positive hedonic tone or quality or that it is above neutral. I would be comparing it to other experiences that I often have, which have more negative aspects, such as feelings of discomfort. When I carefully consider my experiences, I cannot detect that I have ever experienced anything that

[17] Similar ideas can, e.g. be found in Schopenhauer's and Epicurus' writings and Fehige (1998). For a related recent text, see Gloor (2017).

I would say is or warrants being called positive, on the plus side, above neutral or the like. This includes what are commonly considered peak events in life, such as major accomplishments. At such times, my main feeling has been relief, sometimes combined with excitement about what I will do in the future, but the feeling has not had a positive quality (being excited need not feel positive). In contrast, I often have decidedly negative experiences. Actually, the phrases 'negative well-being' and 'negative experiences' are unfortunate because if something is negative, it sounds as if there is a positive counterpart. Better names may be 'problematic moments in life' and 'problematic experiences', because unproblematic, which seems to be the opposite of problematic, does not imply positive.[18]

The final critical question is whether others who claim to have experiences with a positive hedonic tone are mistaken. They may not be introspecting accurately, or they may misremember, but I do not rely on that. Rather, I would say that we can choose how to label an experience and whether to say that it has a positive hedonic tone. We disagree about how to categorise the world and what it is appropriate to call an experience.[19] Here we, unfortunately, seem to reach fundamental judgments or opinions, and I am unsure what more can be said.

4. Is killing everyone more likely to become optimal from a negative or a traditional utilitarian perspective?

Another reply to elimination arguments is that the purportedly wrong acts are unlikely to be optimal in real life. That is, one could argue that it is unlikely that these specific acts will become available and result in the smallest sum of suffering or the greatest sum of well-being if one is defending negative or traditional utilitarianism, respectively. Whether this is a more successful defence of traditional utilitarianism than of negative utilitarianism depends on whether it is more likely from a negative utilitarian perspective than from a traditional utilitarian perspective that a real-world situation will occur in which it is optimal to try to kill everyone. By 'occur', I mean that at some point in time, at least one agent is in a situation wherein the described act is both available and optimal according to the form of utilitarianism under consideration.

It seems roughly equally likely that killing everyone without replacement will become optimal in the real world from a negative utilitarian

[18] I am here inspired by Gloor (2017).

[19] I here draw on related ideas found in Tomasik's work, e.g., his (2017e).

perspective as it is that killing everyone and replacing us with agents with greater well-being will become optimal from a traditional utilitarian perspective. My aim in this section, however, is not to settle the issue, but to briefly argue the following: There are, concerning both negative and traditional utilitarianism, many considerations for and against the plausibility that killing everyone will ever become optimal for an agent in real life. Someone who argues that traditional utilitarianism is more plausible than negative utilitarianism because negative utilitarianism more probably implies that killing everyone will become optimal in real life needs to explain why that is so because it is not obvious.

A first consideration is that in the real world, there are strong tactical reasons from both a negative and a traditional perspective to compromise and accommodate others' wishes, partly to increase the chances that one at least accomplishes one's most important goals (Tomasik 2016b, sec. Why we should remain cooperative). This speaks against killing everyone as a plausible optimal action in real life from either perspective.

Several considerations related to wild animals, evolution and space weigh against that a real-world situation will occur in which negative utilitarianism implies that it is optimal to kill all humans or all sentient beings on Earth. If merely all humans died, there would be room for more suffering wild animals (Tomasik 2017d), and humans would no longer be able to reduce wild-animal suffering, which we may do if we survive (Vinding 2015). Even if all sentient beings on Earth died, beings that suffer could still evolve again on Earth (Acton and Watkins 1963, 96; J. J. C. Smart 1989, 44). Also, if humans survive, we may reduce suffering in other parts of the universe (Pearce 1995, chap 4, objection 32; 2013), or, at least, if we spread through space, it may result in less suffering than if other spacefaring aliens do so instead (Tomasik 2016b). Similarly, if all humans or all sentient beings on Earth were killed, a new spacefaring civilisation may eventually develop on Earth, and if it were to colonise space, it is an open question whether it would result in more suffering than if we were to (Tomasik 2017b). Perhaps most exotically, if we are not killed, humans or our descendants may reduce the number of universes that come into existence naturally, given, for example, multiverse scenarios (Tomasik 2017c). There are also counter-considerations. Human extinction would entail that we would not multiply suffering beyond Earth by colonising space or, more speculatively, by creating new universes (Tomasik 2016b, 2016a, 2017c). At least the killing of all humans, and more so the killing of all sentient life on Earth, would presumably reduce the likelihood of colonisation or the creation of new universes, because a new civilisation that would be able and willing to colonise space or create new universes might not have time to evolve before Earth becomes uninhabitable to such life forms.

One argument for that killing everyone on Earth is unlikely to be optimal assumes a non-causal decision theory and goes as follows: When calculating the expected value of an available act one should account for that one's choice would amount to one data point about how relevantly similar agents may act in relevantly similar situations. If the act under consideration is to kill all humans or all sentient beings on Earth, one should thus take into account that if one chooses to kill, it should increase one's estimate of the likelihood that other relevantly similar agents also act in ways that are disapproved of. This can reduce the expected value of killing because it can increase one's subjective likelihood that others act in ways in which one disapproves oneself. One's act is only one data point, so the update to one's subjective likelihood that others will behave similarly may be modest – but the effect on one's calculation of expected value can be big if sufficiently many relevantly similar agents make choices with sufficiently high stakes from one's perspective. It is sufficiently likely that there are enough such agents because we probably live in a very large (maybe infinite) universe, or maybe even in a multiverse. Thus, killing everyone on Earth is unlikely to be optimal.[20] This argument can, as far as I can see, be used to defend negative and traditional utilitarianism equally well against elimination arguments.

Another set of considerations concern how feasible it would be to kill everyone with and without replacement. A common view seems to be that killing everyone without replacement is more realistic than killing everyone and replacing us with beings with more aggregate well-being. But there are reasons to doubt this view. One reason is that one of the most feasible ways to cause human extinction is by artificial intelligence, which may lead to the creation of many more sentient beings. Someone who tried to kill everyone without replacement would face many obstacles. For example, a nuclear war would seemingly not lead to human extinction, and cobalt bombs do not seem to be the doomsday machines they are sometimes made out to be (cf. Martin 1982; Ball 2006, 2; Robock 2010, 424; Geist 2016, 239–41; Sandberg n.d.). Using pathogens appears more feasible, but still extremely difficult.[21] Another doomsday scenario involves runaway self-replicating nanorobots – so-called 'grey goo' – killing all humans or even consuming the biosphere; however, according to the

[20] I have formulated the argument based on ideas in Oesterheld (2017).

[21] For obstacles to the use of pathogens, see Zilinskas (2001), Shea and Gottron (2004) and Ben Ouagrham-Gormley (2014).

Center for Responsible Nanotechnology (2003), 'goo would be extremely difficult to design and build'.

What about death by artificial intelligence (AI) or artificial general intelligence (AGI)? Roughly speaking, AGI refers to AI with at least human-level intelligence in a wide range of areas. According to Tomasik (2017b), 'the only known technological development that is highly likely to cause all-out human extinction is AGI'. However, extinction by AGI differs crucially from the aforementioned extinction scenarios in that it carries a higher likelihood of humans being replaced by vastly more sentient beings beyond Earth (Tomasik 2017b). One reason to believe this is that an AGI might expand beyond Earth to acquire more resources for pursuing its goals, creating more sentient beings as it expands (see Omohundro 2008, sec. 6). This risk of an astronomical increase in suffering suggests that killing everyone or attempting to do so, using AGI will not become optimal from a negative utilitarian perspective.[22] Let us turn to the traditional utilitarian perspective. Why may an AGI with something like traditional utilitarian values kill all humans or all sentient beings on Earth? An AGI may focus on cosmic stakes, essentially ignore us, and run us over or use up resources so that we starve, in pursuit of its more important goals. Alternatively, we could be killed and used as raw material or fuel for various purposes, including colonising space faster or creating beings that produce positive well-being more effectively. Or we could be killed to prevent us from causing problems for the AGI's plans. That said, developing or unleashing an AGI that will kill everyone involves the risk that something might fail, which could prevent vast amounts of well-being from being created beyond Earth and potentially even cause vast amounts of suffering to be created instead. It is plausible that a careful, peaceful approach to AGI will be optimal from a traditional utilitarian perspective. One can still argue, however, that in some future scenarios, an agent will be sufficiently confident that an AGI will act in line with certain values, and it will become optimal for that agent (based either on traditional or negative utilitarianism) to cause an AGI to kill everyone on Earth. One might make the case that such a scenario is more or less likely to occur from a negative or traditional utilitarian perspective, but this would require a more detailed analysis of AI scenarios.

Finally, in defence of negative utilitarianism, one can argue that it is less likely to become optimal from a negative utilitarian perspective to attempt to use, say, pathogens to kill everyone than it is from a traditional utilitarian

[22] This point has essentially been made by at least Tomasik (2017a, sec. Would a human-inspired AI or rogue AI cause more suffering?, 2018).

perspective to attempt to use an AGI to kill everyone in order to increase the amount of positive well-being (the aim can, but need not, be to also reduce negative well-being). Developing pathogens seems more suspicious than developing an AGI because the AGI will more likely have ostensibly beneficial features. Thus, the argument goes, the development and use of the AGI involve less risk of getting caught and stopped.

5. Indirect act-utilitarianism

The final type of reply to elimination arguments that I will consider is the appeal to indirect act-utilitarianism. Indirect act-utilitarianism combines the act-utilitarian criterion for the rightness of acts with the idea that as a practical matter we tend to act rightly (take the actions with the best results) if we, in general, do not think in act-utilitarian terms when conducting our lives and deciding what to do in particular cases. Instead, to indirectly optimise the results of our actions, we should, for example, internalise deontological moral rules and develop various character traits.[23] An indirect act-utilitarian might argue that, in the real world, a moral person will, for practical reasons, develop dispositions that prevent her from killing everyone (cf. Hare 1981, 135).

This appeal to indirect act-utilitarianism is not a satisfactory reply to elimination arguments. First, it is compatible with the claim that killing everyone would, in some cases, be right. The reason is simply that indirect act-utilitarianism includes the act-utilitarian criterion for the rightness of acts. So, regardless of whether someone with the optimal dispositions *would* or *would not* kill everyone, the specific act of killing everyone could still be optimal and hence right. Second, even if it were optimal for people *generally* to have dispositions that would prevent them from killing, what dispositions and rules one should adopt to indirectly optimise the results of one's acts vary by person and time. It is plausible that, if killing everyone would be the right act, then it would likely be optimal for at least someone – say, a president, dictator or corporate leader – to be prepared to kill everyone in special circumstances when vast amounts of well-being are at stake. Third, if the indirect act-utilitarianism includes the idea that some people should sometimes make act-utilitarian calculations when deciding what to do in particular cases, then the huge-stakes choices in the elimination arguments are strong candidates for situations wherein calculating would indeed be prescribed.

[23] For more on indirect act-utilitarianism, see Wiland (2017).

Finally, regardless of whether the appeal to indirect act-utilitarianism is a convincing response, similar considerations come into play whether we investigate the implications of negative or traditional utilitarianism. For instance, a traditional utilitarian may argue that others get subtle indications of someone's dispositions when interacting with them. Being open to killing everyone would not be optimal because others would be able to tell that something about the person is awry, which would hamper the person's ability to bring about the best outcomes. But a negative utilitarian can make the same argument. Therefore, the appeal to indirect act-utilitarianism leaves traditional utilitarianism no better off than negative utilitarianism.

6. Conclusions and future research

World destruction- or elimination arguments exist against both traditional and negative utilitarianism; proponents of both theories have several replies available. I have not found any of these replies more convincing when offered in defence of traditional utilitarianism versus in defence of negative utilitarianism. Hence, I find that it is a mistake to consider negative utilitarianism more vulnerable to elimination arguments than is traditional utilitarianism. As things stand, elimination arguments are not a reason to reject negative utilitarianism in favour of traditional utilitarianism. I have compared negative utilitarianism with traditional utilitarianism, but a comparison of negative utilitarianism with, say, prioritarianism with regard to vulnerability to elimination arguments would have a similar structure, and I expect the result to be similar. I, therefore, conclude that those who argue against negative utilitarianism in favour of such other consequentialist views need to rely on other arguments or explain why their theory is less vulnerable to elimination arguments than negative utilitarianism.

I have focused on elimination arguments, but one can make similar analyses of other common objections to negative utilitarianism, especially when the objection is meant to show that some other form of consequentialism is more plausible than negative utilitarianism. For example, one common objection to negative utilitarianism is that it purportedly implies that one has no obligation to raise the positive well-being of many individuals, prevent a decrease in their positive well-being, or bring into existence many new beings with positive well-being, even if the cost of doing so would be zero or trivial (Griffin 1979, 48; Bykvist 2010, 62; Hurka 2010, 200). Analysing this objection along the lines of this paper would likely show that a negative utilitarian has several good replies available. A reply that I endorsed above is that there is no positive well-being. Other replies can build on the idea that if the objector wants to avoid unrealistic counterexamples to her morality, counterexamples to negative utilitarianism need to be realistic. The negative utilitarian can then reply that, in real life, one cannot increase or prevent a decrease in positive well-being at no or trivial cost, because the cost is that the agent instead could have tried to prevent severe suffering. Moreover, in the real world, negative utilitarianism may imply that one should, in general, increase others' positive well-being if there is such a thing and if it could be done at no or trivial cost because those who have higher positive well-being tend to suffer less. Finally, the negative utilitarian can argue that it is optimal to be cooperative and take into account what others value, and because many others care about positive well-being, one should increase it if one can do so at low cost.[24]

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Simon Knutsson http://orcid.org/0000-0002-0319-8273

References

Acton, H. B., and J. W. N. Watkins. 1963. "Symposium: Negative Utilitarianism." *Aristotelian Society Supplementary Volume* 37: 83–114.

Animal Ethics. 2016. "The Situation of Animals in the Wild." https://www.animal-ethics.org/wild-animal-suffering-section/wild-animals/.

Annas, Julia. 1993. *The Morality of Happiness*. New York: Oxford University Press.

Arrhenius, Gustaf, and Krister Bykvist. 1995. *Future Generations and Interpersonal Compensations: Moral Aspects of Energy Use*. Uppsala: Uppsala Prints and Preprints in Philosophy.

Ball, Desmond. 2006. *The Probabilities of On the Beach: Assessing 'Armageddon Scenarios' in the 21st Century*. Working Paper No. 401. Canberra, A.C.T: Strategic and Defence Studies Centre, The Australian National University.

[24] I am grateful for comments on earlier versions of this paper from Lars Bergström, Erik Carlson, Ruairí Donnelly, Gustav Ekman, Oscar Horta, Jens Johansson, Kaj Sotala, Johannes Treutlein, Torbjörn Tännsjö, and especially Krister Bykvist, Max Daniel, Brian Tomasik and Magnus Vinding. The paper has benefited from correspondence with Tobias Baumann, Dale Jamieson, Caspar Oesterheld, David Pearce and Anders Sandberg. Thanks to Adrian Rorheim for copy editing an earlier version, and to three anonymous reviewers for helpful comments. I did the early stage of the work on the paper when I was employed by the Foundational Research Institute. Views expressed are not necessarily those of the institute.

Baum, Seth D., and Grant S. Wilson. 2013. "The Ethics of Global Catastrophic Risk from Dual-Use Bioengineering." *Ethics in Biology, Engineering and Medicine: An International Journal* 4 (1): 59–72.

Benatar, David. 2017. *The Human Predicament: A Candid Guide to Life's Biggest Questions*. New York: Oxford University Press.

Ben Ouagrham-Gormley, Sonia. 2014. *Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development*. Ithaca: Cornell University Press.

Bunge, Mario. 1989. *Treatise on Basic Philosophy: Volume 8: Ethics: The Good and the Right*. Dordrecht: Springer.

Bykvist, Krister. 2010. *Utilitarianism: A Guide for the Perplexed*. London: Continuum.

Center for Responsible Nanotechnology. 2003. "Nanotechnology: Grey Goo Is a Small Issue." http://crnano.org/BD-Goo.htm.

Cooney, Nick. 2013. *Veganomics: The Surprising Science on Vegetarians, from the Breakfast Table to the Bedroom*. New York: Lantern Books.

Edwards, Rem B. 1986. "The Principle of Utility and Mill's Minimizing Utilitarianism." *Journal of Value Inquiry* 20 (2): 125–136.

Fehige, Christoph. 1998. "A Pareto Principle for Possible People." In *Preferences*, edited by Christoph Fehige, and Ulla Wessels, 508–543. Berlin: De Gruyter.

Geist, Edward Moore. 2016. "Would Russia's Undersea "Doomsday Drone" Carry a Cobalt Bomb?" *Bulletin of the Atomic Scientists* 72 (4): 238–242.

Gloor, Lukas. 2017. "Tranquilism." Foundational Research Institute. https://foundational-research.org/tranquilism/.

Griffin, James. 1979. "Is Unhappiness Morally More Important than Happiness?" *Philosophical Quarterly* 29 (114): 47–55.

Hare, R. M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Clarendon.

Hedenius, Ingemar. 1955. *Fyra dygder*. Stockholm: Albert Bonniers Förlag.

Heyd, David. 1992. *Genethics: Moral Issues in the Creation of People*. Berkeley: University of California Press.

Horta, Oscar. 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.

Hurka, Thomas. 2010. "Asymmetries in Value." *Noûs* 44 (2): 199–223.

Jamieson, Dale. 1984. "Utilitarianism and the Morality of Killing." *Philosophical Studies* 45 (2): 209–221.

Martin, Brian. 1982. "Critique of Nuclear Extinction." *Journal of Peace Research* 19 (4): 287–300.

Matthews, Richard. 2008. *The Absolute Violation: Why Torture Must Be Prohibited*. Montreal and Kingston: McGill-Queen's University Press.

McCloskey, H. J. 1957. "An Examination of Restricted Utilitarianism." *The Philosophical Review* 66 (4): 466–485.

Mukwege, Denis Mukengere, and Cathy Nangini. 2009. "Rape with Extreme Violence: The New Pathology in South Kivu, Democratic Republic of Congo." *PLoS Medicine* 6 (12): e1000204.

Mulgan, Tim. 2007. *Understanding Utilitarianism*. Stocksfield: Acumen.

Ng, Yew-Kwang. 1995. "Towards Welfare Biology: Evolutionary Economics of Animal Consciousness and Suffering." *Biology and Philosophy* 10 (3): 255–285.

Norwood, F. Bailey, and Jayson Lusk. 2011. *Compassion, by the Pound: The Economics of Farm Animal Welfare*. New York: Oxford University Press.

Oesterheld, Caspar. 2017. “Multiverse-Wide Cooperation via Correlated Decision Making.” Foundational Research Institute. https://foundational-research.org/multiverse-wide-cooperation-via-correlated-decision-making/.

Omohundro, Stephen M. 2008. “The Nature of Self-Improving Artificial Intelligence.” https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf.

Ord, Toby. 2013. “Why I’m Not a Negative Utilitarian.” http://www.amirrorclear.net/academic/ideas/negative-utilitarianism/, http://www.webcitation.org/6bmUNUHHW.

Pearce, David. 1995. “The Hedonistic Imperative.” http://www.hedweb.com/.

Pearce, David. 2005. “The Pinprick Argument.” https://www.utilitarianism.com/pinprick-argument.html, http://www.webcitation.org/6nhxs7oNW.

Pearce, David. 2013. “Unsorted Postings.” https://www.hedweb.com/social-media/pre2014.html, http://www.webcitation.org/6mRy7BQrW.

Peters, Edward. 1985. *Torture*. New York: Blackwell.

Pluhar, Evelyn. 1990. “Utilitarian Killing, Replacement, and Rights.” *Journal of Agricultural Ethics* 3 (2): 147–171.

Robertson, Michael, Kirsty Morris, and Garry Walter. 2007. “Overview of Psychiatric Ethics V: Utilitarianism and the Ethics of Duty.” *Australasian Psychiatry* 15 (5): 402–410.

Robock, Alan. 2010. “Nuclear Winter.” *Wiley Interdisciplinary Reviews: Climate Change* 1 (3): 418–427.

Sandberg, Anders. n.d. “Nuclear Holocaust.” Unpublished manuscript.

Shea, Dana A., and Frank Gottron. 2004. “Small-Scale Terrorist Attacks Using Chemical and Biological Agents: An Assessment Framework and Preliminary Comparisons.” Congressional Research Service. http://www.fas.org/irp/crs/RL32391.pdf.

Smart, R. N. 1958. “Negative Utilitarianism.” *Mind* 67 (268): 542–543.

Smart, J. J. C. 1973. “An Outline of a System of Utilitarian Ethics.” In *Utilitarianism: For and Against*, edited by J. J. C. Smart, and Bernard Williams, 3–74. London: Cambridge University Press.

Smart, J. J. C. 1989. “Negative Utilitarianism.” In *Freedom and Rationality*, edited by Fred D’Agostino, and I. C. Jarvie, 35–46. Dordrecht: Kluwer Academic.

Sprigge, T. L. S. 1965. “A Utilitarian Reply to Dr. McCloskey.” *Inquiry* 8 (1–4): 264–291.

Tännsjö, Torbjörn. 1997. “Doom Soon?” *Inquiry: An Interdisciplinary Journal of Philosophy* 40 (2): 243–252.

Tännsjö, Torbjörn. 2015. “Utilitarianism or Prioritarianism?” *Utilitas* 27 (2): 240–250.

Tännsjö, Torbjörn. 2016. “Därför blir universum en bättre plats när människan är borta.” *Dagens Nyheter*, May 8, 2016. https://www.dn.se/arkiv/kultur/darfor-blir-universum-en-battre-plats-nar-manniskan-ar-borta/.

Thomson, Judith Jarvis. 1976. “Killing, Letting Die, and the Trolley Problem.” *The Monist* 59 (2): 204–217.

Tomasik, Brian. 2016a. “Applied Welfare Biology and Why Wild-Animal Advocates Should Focus on Not Spreading Nature.” Essays on Reducing Suffering. http://reducing-suffering.org/applied-welfare-biology-wild-animal-advocates-focus-spreading-nature/.

Tomasik, Brian. 2016b. “Risks of Astronomical Future Suffering.” Foundational Research Institute. https://foundational-research.org/risks-of-astronomical-future-suffering/.

Tomasik, Brian. 2017a. “Artificial Intelligence and Its Implications for Future Suffering.” Foundational Research Institute. https://foundational-research.org/artificial-intelligence-and-its-implications-for-future-suffering.

Tomasik, Brian. 2017b. “How Would Catastrophic Risks Affect Prospects for Compromise?” Foundational Research Institute. https://foundational-research.org/how-would-catastrophic-risks-affect-prospects-for-compromise/.

Tomasik, Brian. 2017c. “Lab Universes: Creating Infinite Suffering.” Essays on Reducing Suffering. http://reducing-suffering.org/lab-universes-creating-infinite-suffering/.

Tomasik, Brian. 2017d. “Strategic Considerations for Moral Antinatalists.” Essays on Reducing Suffering. http://reducing-suffering.org/strategic-considerations-moral-antinatalists/.

Tomasik, Brian. 2017e. “Which Computations Do I Care About?” Essays on Reducing Suffering. https://reducing-suffering.org/which-computations-do-i-care-about/.

Tomasik, Brian. 2018. “Summary of My Beliefs and Values on Big Questions.” Essays on Reducing Suffering. https://reducing-suffering.org/summary-beliefs-values-big-questions/.

Vinding, Magnus. 2015. *Anti-Natalism and the Future of Suffering: Why Negative Utilitarians Should Not Aim for Extinction*. Smashwords.

Wiland, Eric. 2017. “Indirect Utilitarianism.” In *Bloomsbury Encyclopedia of Utilitarianism*, edited by James E. Crimmins, 269–272. London: Bloomsbury.

Zilinskas, Raymond A. 2001. “Possible Terrorist Use of Modern Biotechnology Techniques.” In *Emerging Technologies: Recommendations for Counter-Terrorism*, edited by Joseph Rosen, and Charles Lucey, 106–121. Hanover, NH: Institute for Security Technology Studies, Dartmouth College. http://www.dartmouth.edu/~engs05/md/whitepapers/Emerging_Tech/ETech.pdf.