Superintelligence as a Cause or Cure for Risks of Astronomical Suffering

Kaj Sotala
Foundational Research Institute
kaj.sotala@foundational-research.org, foundational-research.org

Lukas Gloor
Foundational Research Institute
lukas.gloor@foundational-research.org, foundational-research.org

Keywords: existential risk, suffering risk, superintelligence, AI alignment problem, suffering-focused ethics, population ethics

Received: 31 August 2017

Abstract: Discussions about the possible consequences of creating superintelligence have included the possibility of existential risk, often understood mainly as the risk of human extinction. We argue that suffering risks (s-risks), where an adverse outcome would bring about severe suffering on an astronomical scale, are risks of a comparable severity and probability as risks of extinction. Preventing them is the common interest of many different value systems. Furthermore, we argue that in the same way as superintelligent AI both contributes to existential risk but can also help prevent it, superintelligent AI can be both the cause of suffering risks and a way to prevent them from being realized. Some types of work aimed at making superintelligent AI safe will also help prevent suffering risks, and there may also be a class of safeguards for AI that helps specifically against s-risks.

1 Introduction

Work discussing the possible consequences of creating superintelligent AI (Yudkowsky 2008, Bostrom 2014, Sotala & Yampolskiy 2015) has discussed it as a possible *existential risk*: a risk "where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" (Bostrom 2002, 2013).

The previous work has mostly[1] considered the worst-case outcome to be the possibility of human extinction by a superintelligent AI (henceforth *superintelligence*) that is indifferent to humanity's survival and values. However, it is often thought that for an individual, there exist "fates worse than death"; analogously, for civilizations there may exist fates worse than extinction, such as survival in conditions in which most people will experience enormous suffering for most of their lives.

Even if such extreme outcomes would be avoided, the known universe may eventually be populated by vast amounts of minds: published estimates include the possibility of 10^25 minds supported by a single star (Bostrom 2003a), with humanity having the potential to eventually colonize tens of millions of galaxies (Armstrong & Sandberg 2013). While this could enable an enormous number of meaningful lives to be lived, if even a small fraction of these lives were to exist in hellish circumstances, the amount of suffering would be vastly greater than that produced by all the atrocities, abuses, and natural causes in Earth's history so far.

We term the possibility of such outcomes a *suffering risk:*

> Suffering risk (s-risk): One where an adverse outcome would bring about severe suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far.

In order for potential risks—including s-risks— to merit work on them, three conditions must be met. First, the outcome of the risk must be sufficiently severe to merit attention. Second, the risk must have some reasonable probability of being realized. Third, there must be some way for risk-avoidance work to reduce either the probability or severity of an adverse outcome.

In this paper, we will argue that suffering risks meet all three criteria, and that s-risk avoidance work is thus of a comparable magnitude in importance as work on risks from extinction. Section 2 seeks to establish the severity

[1] Bostrom (2014) focuses mainly on the risk of extinction, but also devotes some discussion to other negative outcomes such as "mindcrime" (see section 5).

of s-risks. There, we will argue that there are classes of suffering-related adverse outcomes that many value systems would consider to be equally or even more severe than extinction. Additionally, we will define a class of less severe suffering outcomes which many value systems would consider important to avoid, albeit not as important as avoiding extinction. Section 3 looks at suffering risks from the view of several different value systems, and discusses how much they would prioritize avoiding different suffering outcomes. Next, we will argue that there is a reasonable probability for a number of different suffering risks to be realized. Our discussion is organized according to the relationship that superintelligent AIs have to suffering risks: section 4 covers risks that may be prevented by a superintelligence, and section 5 covers risks that may be realized by one.[2] Section 6 discusses how it might be possible to work on suffering risks.

2 Suffering risks as risks of extreme severity

As already noted, the main focus in discussion of risks from superintelligent AI has been either literal extinction, with the AI killing humans as a side-effect of pursuing some other goal (Yudkowsky 2008), or a *value extinction*. In value extinction, some form of humanity may survive, but the future is controlled by an AI operating according to values all current-day humans would consider worthless (Yudkowsky 2011). In either scenario, it is thought that the resulting future would have no value.

In this section, we will argue that besides futures that have *no value*, according to many different value systems it is possible to have futures with *negative value*. These would count as the worst category of existential risks. In addition, there are adverse outcomes of a lesser severity, which depending on one's value systems may not necessarily count as worse than extinction. Regardless, making these outcomes less likely is a high priority and a common interest of many different value systems.

Bostrom (2002) frames his definition of extinction risks with a discussion which characterizes a single person's death as being a risk of terminal intensity and personal scope, with existential risks being risks of terminal intensity and *global* scope—one person's death versus the death of all humans. However, it is commonly thought that there are "fates worse than death": at one extreme, being tortured for an extended time (with no chance of rescue), and then killed.

As less extreme examples, various negative health conditions are often considered worse than death (Rubin, Buehler & Halpern 2016; Sayah et al. 2015; Ditto et al., 1996): for example, among hospitalized patients with severe illness, a majority of respondents considered bowel and bladder incontinence, relying on a feeding tube to live, and being unable to get up from bed, to be conditions that were worse than death (Rubin, Buehler & Halpern 2016). While these are prospective evaluations rather than what people have actually experienced, several countries have laws allowing for voluntary euthanasia, which people with various adverse conditions have chosen rather than go on living. This may considered an empirical confirmation of some states of life being worse than death, at least as judged by the people who choose to die.

The notion of fates worse than death suggests the existence of a "hellish" severity that is one level worse than "terminal", and which might affect civilizations as well as individuals. Bostrom (2013) seems to acknowledge this by including "hellish" as a possible severity in the corresponding chart, but does not place any concrete outcomes under the hellish severity, implying that risks of extinction are still the worst outcomes. Yet there seem to be plausible paths to civilization-wide hell outcomes as well (Figure 1), which we will discuss in sections 4 and 5.

Global	Thinning of the ozone layer	**Extinction risks**	**Global hellscape**
Personal	Car is stolen	Death	Extended torture followed by death
	Endurable	**Terminal**	**Hellish**

Figure 1: The worst suffering risks are ones that affect everyone and subject people to hellish conditions.

In order to qualify as equally bad or worse than extinction, suffering risks do not necessarily need to affect every single member of humanity. For example, consider a simplified ethical calculus where someone may have a predominantly happy life (+1), never exist (0), or have a predominantly unhappy life (-1). As long as the people having predominantly unhappy lives outnumber the people having predominantly happy lives, under this calculus such an outcome would be considered worse than nobody existing in the first place. We will call this scenario a **net suffering outcome.**[3]

[2] Superintelligent AIs being in a special position where they might either enable or prevent suffering risks, is similar to the way in which they are in a special position to make risks of extinction both more or less likely (Yudkowsky 2008).

[3] "Net" should be considered equivalent to Bostrom's "global", but we have chosen a different name to avoid

This outcome might be considered justifiable if we assumed that, given enough time, the people living happy lives would eventually outnumber the people living unhappy lives. Most value systems would then still consider a net suffering outcome worth avoiding, but they might consider it an acceptable cost for an even larger amount of future happy lives.

On the other hand it is also possible that the world could become locked into conditions in which the balance would remain negative even when considering all the lives that will ever live: things would never get better. We will call this a **pan-generational net suffering outcome.**

In addition to net and pan-generational net suffering outcomes, we will consider a third category. In these outcomes, serious suffering may be limited to only a fraction of the population, but the overall population at some given time[4] is still large enough that even this small fraction accounts for many times more suffering than has existed in the history of the Earth. We will call these **astronomical suffering outcomes.**

Types of suffering outcomes	
Astronomical suffering outcome	At some point in time, a fraction of the population experiences hellish suffering, enough to overall constitute an astronomical amount that overwhelms all the suffering in Earth's history.
Net suffering outcome	At some point in time, there are more people experiencing lives filled predominantly with suffering than there are people experiencing lives filled predominantly with happiness.
Pan-generational net suffering outcome	When summed over all the people that will ever live, there are more people experiencing lives filled predominantly with suffering than there are people experiencing lives filled predominantly with happiness.

Figure 2: types of possible suffering outcomes. An outcome may count as one or several of the categories in this table.

Any value system that puts weight on preventing suffering implies at least some interest in preventing suffering risks. Additionally, as we will discuss below, even value systems which do not care about suffering *directly* may still have an interest in preventing suffering risks.

We expect these claims to be relatively uncontroversial. A more complicated question is that of *tradeoffs*: what should one do if some interventions increase the risk of extinction but make suffering risks less likely, or vice versa? As we will discuss below, if forced to choose between these two, different value systems will differ in which of the interventions they favor. In such a case, rather than to risk conflict between value systems, a better alternative would be to attempt to identify interventions that do not involve such a tradeoff. If there were interventions that reduced the risk of extinction without increasing the risk of astronomical suffering, or decreased the risk of astronomical suffering without increasing the risk of extinction, or decreased both, then it would be in everyone's interest to agree to jointly focus on these three classes of interventions.

3 Suffering risks from the perspective of different value systems

We will now take a brief look at different value systems and their stance on suffering risks, as well as their stance on the related tradeoffs.

Classical utilitarianism. All else being equal, classical utilitarians would prefer a universe in which there were many happy lives and no suffering. However, a noteworthy feature about classical utilitarianism (as well as some other aggregative theories) is that it considers very good and very bad scenarios to be symmetrical—that is, a scenario with 10^20 humans living happy lives may be considered equally good, as a scenario with 10^20 humans living miserable lives is considered bad.

Thus, people following classical utilitarianism or some other aggregative theory may find compelling the argument (Bostrom 2003a) that an uncolonized universe represents a massive waste of potential value, and be willing to risk—or even accept—astronomical numbers of suffering individuals if that was an unavoidable cost to creating even larger numbers of happiness. Thus, classical utilitarianism would consider astronomical and net suffering outcomes something to avoid but possibly acceptable, and pan-generational net suffering outcomes as something to avoid under all circumstances.

Other aggregative theories. Any moral theory that is not explicitly utilitarian, yet still has an aggregative component that disvalues suffering, would consider suffering risks as something to avoid. Additionally, for moral theories that value things other than just pleasure and suffering—such as preference satisfaction, some broader notion of "human flourishing", objective list theories—hellscape scenarios would likely also threaten the satisfaction of many of the things that these theories valued. For example, minds experiencing enormous suffering are probably not flourishing, are likely to have

giving the impression that the outcome would necessarily be limited to only one planet.

[4] One could also consider the category of pan-generational astronomical suffering outcomes, but restricting ourselves into just three categories is sufficient for our current discussion.

unsatisfied preferences, and probably do not have many of the things considered valuable in objective list theories.

Similarly to classical utilitarianism, many aggregative theories could be willing to risk or even accept astronomical and civilization-wide suffering outcomes as a necessary evil, yet wish to avoid pan-generational net suffering outcomes. At the same time, many aggregative theories might incorporate some suffering-focused intuition (discussed below) that cause them to put more weight on the avoidance of suffering than the creation of other valuable things. Depending on the circumstances, this might cause them to reject the kind of reasoning that suggested that suffering outcomes could be an acceptable cost.

Rights-based theories. Rights-based theories would consider suffering risks a bad thing *directly* to the extent that they held that people—or animals (Regan 1980)—had a right to be treated well avoid unnecessary suffering. They could also consider suffering risks *indirectly* bad, if the suffering was caused by conditions which violated some other right or severely constrained someone's capabilities (Nussbaum 1997, p. 287). For example, a right to meaningful autonomy could be violated if a mind was subjected to enormous suffering and had no meaningful option to escape it.

General suffering-focused intuitions. There are various moral views and principles which could fit many different value systems, all of which would imply that suffering risks were something important to avoid and which might cause one to weigh the avoidance of suffering more strongly than the creation of happiness:

1. *Prioritarianism.* Prioritarianism is the position that the worse off an individual is, the more morally valuable it is to make that individual better off (Parfit 1991). That is, if one person is living in hellish conditions and another is well-off, then making the former person slightly better off is more valuable than improving the life of the well-off person by the same amount. A stance of "astronomical prioritarianism" that considers all minds across the universe, and prioritizes improving the worst ones sufficiently strongly, pushes in the direction of mainly improving the lives of those that would be worst off and thus avoiding suffering risks. If a suffering outcome does manifest itself, prioritarianism would prioritize bringing it to an end, over creating additional well-off lives or further helping those who are already well off. Prioritarianism may imply focusing particularly on risks from future technologies, as these may enable the creation of mind states that are worse than the current biopsychological limits.

Besides prioritarianism, the following three intuitions (Gloor & Mannino 2016) would also prioritize the avoidance of suffering risks:[5]

2. *Making people happy, not happy people.*[6] An intuition which is present in preference-based views such as antifrustrationism (Fehige 1998), antinatalism (Benatar 2008), as well as the "moral ledger" analogy (Singer 1993) and prior-existence utilitarianism (Singer 1993), is that it is more important to make existing people better off than it is to create new happy beings.[7] For example, given the choice between helping a million currently-existing people who are in pain and bringing ten million new people into existence, this view holds that it is more important to help the existing people, even if the ten million new people would end up living happy lives.

A part of this view is the notion that it is not intrinsically bad to never be created, whereas it is intrinsically bad to exist and be badly off, or to be killed against one's wishes once one does exist. If one accepts this position, then one could still want to avoid extinction—or at least the death of currently-living humans—but the promise of astronomical numbers of happy lives being created (Bostrom 2003a) would not be seen as particularly compelling, whereas the possible creation of astronomical numbers of lives experiencing suffering could be seen as a major thing to avoid.

3. *Torture-level suffering cannot be counterbalanced.* This intuition is present in the widespread notion that minor pains cannot be aggregated to become worse than an instant of torture (Rachels 1998), in threshold negative utilitarianism (Ord 2013), philosophical fictional works such as *The Ones Who Walk Away From Omelas* (LeGuin 1973), and it may contribute to the absolute prohibitions against torture in some deontological moralities. Pearce (1995) expresses a

[5] One might naturally also have various intuitions that point in the opposite direction, that is, of not prioritizing suffering risks. We will not survey these, as our intent in this section is merely to establish that many would consider suffering risks as important to avoid, without claiming that this would be the *only* plausible view to hold.

[6] The name of this intuition is a paraphrase of Narveson (1973), "We are in favor of making people happy, but neutral about making happy people."

[7] Moral views that attempt to incorporate this intuition by treating the creation of new people as morally neutral (e.g. Singer's "prior-existence" criterion) suffer from what Greaves (2017) calls a "remarkabl[e] difficult[y] to formulate any remotely acceptable axiology that captures this idea of 'neutrality'". The views by Benatar and Fehige avoid this problem, but they imply a more extreme position where adding new lives is neutral only in a best-case scenario where they contain no suffering or frustrated preferences.

form of it when he writes, “No amount of happiness or fun enjoyed by some organisms can notionally justify the indescribable horrors of Auschwitz”.

4. *Happiness as the absence of suffering*. A view present in Epicureanism, as well as many non-Western traditions such as Buddhism, is that of happiness as the absence of suffering. Under this view, when we are not experiencing states of pleasure, we begin to crave pleasure, and this craving constitutes suffering. Gloor (2017) writes:

> Uncomfortable pressure in one’s shoes, thirst, hunger, headaches, boredom, itches, non-effortless work, worries, longing for better times. When our brain is flooded with pleasure, we temporarily become unaware of all the negative ingredients of our stream of consciousness, and they thus cease to exist. Pleasure is the typical way in which our minds experience temporary freedom from suffering. This may contribute to the view that pleasure is the symmetrical counterpart to suffering, and that pleasure is in itself valuable and important to bring about. However, there are also (contingently rare) mental states devoid of anything bothersome that are not commonly described as (intensely) pleasurable, examples being flow states or states of meditative tranquility. Felt from the inside, tranquility is perfect in that it is untroubled by any aversive components, untroubled by any cravings for more pleasure. Likewise, a state of flow as it may be experienced during stimulating work, when listening to music or when playing video games, where tasks are being completed on auto-pilot with time flying and us having a low sense of self, also has this same quality of being experienced as completely problem-free. Such states—let us call them states of contentment—may not commonly be described as (intensely) pleasurable, but following philosophical traditions in both Buddhism and Epicureanism, these states, too, deserve to be considered states of happiness.

Under this view, happiness and pleasure are not intrinsically good, but rather *instrumentally* good in that pleasure takes our focus away from suffering and thus helps us avoid it. Creating additional happiness, then, has no intrinsic value if that creation does not help avoid suffering.

4 Suffering outcomes that could be prevented by a superintelligence

In the previous section, we argued that nearly all plausible value systems will want to avoid suffering risks and that for many value systems, suffering risks are some of the worst possible outcomes and thus some of the most important to avoid. However, whether this also makes suffering risks the type of risk that is the most important to *focus on*, also depends on how probable suffering risks are. If they seem exceedingly unlikely, then there is little reason to care about them.

In this and the next section, we will discuss reasons for believing that there are various suffering outcomes that might realize themselves. We begin by considering outcomes that occur naturally but could be prevented by a superintelligence. In the next section, we will consider suffering outcomes that could be caused by a superintelligence.

A superintelligence could prevent almost any outcome if it established itself a singleton, "a world order in which there is a single decision-making agency at the highest level" (Bostrom 2005). Although a superintelligence is not the only way by which a singleton might be formed, alternative ways—such as a world government or convergent evolution leading everyone to adopt the same values and goals (Bostrom 2005)—do not seem particularly likely to happen soon. Once a superintelligence had established itself as a singleton, depending on its values it might choose to take actions that prevented suffering outcomes from arising.

4.1 Are suffering outcomes likely?

Bostrom (2003a) argues that a technologically mature civilization capable of space colonization on a large scale "would likely also have the ability to establish at least the minimally favorable conditions required for future lives to be worth living", and that it could thus be assumed that all of these lives would be worth living. Moreover, we can reasonably assume that outcomes that are *optimized* for everything that is valuable are more likely than outcomes optimized for things that are disvaluable. While people want the future to be valuable both for altruistic and self-oriented reasons, no one intrinsically wants things to go badly.

However, Bostrom has himself later argued that technological advancement combined with evolutionary forces could "lead to the gradual elimination of all forms of being worth caring about" (Bostrom 2005), admitting the possibility that there could be technologically advanced civilizations with very little of anything that we would consider valuable. The technological potential to create a civilization that had positive value does not automatically translate to that potential being used, so a very advanced civilization could still be one of no value or even negative value.

Examples of technology’s potential being unevenly applied can be found throughout history. Wealth remains unevenly distributed today, with an estimated 795 million people suffering from hunger even as one third of all produced food goes to waste (World Food Programme, 2017). Technological advancement has helped prevent many sources of suffering, but it has also created new ones, such as factory-farming practices under which large numbers of animals are maltreated in ways which maximize their production: in 2012, the number of animals slaughtered for food was estimated at 68 billion worldwide (Food and Agriculture Organization of the United Nations 2012). Industrialization has also contributed to anthropogenic climate change, which may lead to considerable global destruction. Earlier in history, advances in seafaring enabled the transatlantic slave trade, with close to 12 million Africans being sent in ships to live in slavery (Manning 1992).

Technological advancement does not automatically lead to positive results (Häggström 2016). Persson & Savulescu (2012) argue that human tendencies such as "the bias towards the near future, our numbness to the suffering of great numbers, and our weak sense of responsibility for our omissions and collective contributions", which are a result of the environment humanity evolved in, are no longer sufficient for dealing with novel technological problems such as climate change and it becoming easier for small groups to cause widespread destruction. Supporting this case, Greene (2013) draws on research from moral psychology to argue that morality has evolved to enable mutual cooperation and collaboration within a select group ("us"), and to enable groups to fight off everyone else ("them"). Such an evolved morality is badly equipped to deal with collective action problems requiring global compromises, and also increases the risk of conflict and generally negative-sum dynamics as more different groups get in contact with each other.

As an opposing perspective, West (2017) argues that while people are often willing to engage in cruelty if this is the easiest way of achieving their desires, they are generally "not evil, just lazy". Practices such as factory farming are widespread not because of some deep-seated desire to cause suffering, but rather because they are the most efficient way of producing meat and other animal source foods. If technologies such as growing meat from cell cultures became more efficient than factory farming, then the desire for efficiency could lead to the elimination of suffering. Similarly, industrialization has reduced the demand for slaves and forced labor as machine labor has become more effective. At the same time, West acknowledges that this is not a knockdown argument against the possibility of massive future suffering, and that the desire for efficiency could still lead to suffering outcomes such as simulated game worlds filled with sentient non-player characters (see section on cruelty-enabling technologies below).

Another argument against net suffering outcomes is offered by Shulman (2012), who discusses the possibility of civilizations spending some nontrivial fraction of their resources constructing computing matter that was optimized for producing maximum pleasure per unit of energy, or for producing maximum suffering per unit of energy. Shulman's argument rests on the assumption that value and disvalue are symmetrical with regard to such optimized states. The amount of pleasure or suffering produced this way could come to dominate any hedonistic utilitarian calculus, and even a weak benevolent bias that led to there being more optimized pleasure than optimized suffering could tip the balance in favor of there being more total happiness. Shulman's argument thus suggests that net suffering outcomes could be unlikely unless a (non-compassionate) singleton ensures that no optimized happiness is created. However, the possibility of optimized suffering and the chance of e.g. civilizations intentionally creating it as a way of extorting agents that care about suffering reduction, also makes astronomical suffering outcomes more likely.

4.2 Suffering outcome: dystopian scenarios created by non-value-aligned incentives.

Bostrom (2005, 2014) discusses the possibility of technological development and evolutionary and competitive pressures leading to various scenarios where everything of value has been lost, and where the overall value of the world may even be negative. Considering the possibility of a world where most minds are brain uploads doing constant work, Bostrom (2014) points out that we cannot know for sure that happy minds are the most productive under all conditions: it could turn out that anxious or unhappy minds would be more productive. If this were the case, the resulting outcomes could be dystopian indeed:

> We seldom put forth full effort. When we do, it is sometimes painful. Imagine running on a treadmill at a steep incline—heart pounding, muscles aching, lungs gasping for air. A glance at the timer: your next break, which will also be your death, is due in 49 years, 3 months, 20 days, 4 hours, 56 minutes, and 12 seconds. You wish you had not been born. (Bostrom 2014, p. 201)

As Bostrom (2014) notes, this kind of a scenario is by no means inevitable; Hanson (2016) argues for a more optimistic outcome, where brain emulations still spend most of their time working, but are generally happy. But even Hanson's argument depends on economic pressures and human well-being happening to coincide: absent such a happy coincidence, he offers no argument for believing that the future will indeed be a happy one.

More generally, Alexander (2014) discusses examples such as tragedies of the commons, Malthusian traps, arms races, and races to the bottom as cases where people are forced to choose between sacrificing some of their values and getting outcompeted. Alexander also notes the existence of changes to the world that nearly everyone would agree to be net improvements—such as every country reducing its military by 50%, with the savings going to infrastructure—which nonetheless do not happen because nobody has the incentive to carry them out. As such, even if the prevention of various kinds of suffering outcomes would be in everyone's interest, the world might nonetheless end up in them if the incentives are sufficiently badly aligned and new technologies enable their creation.

An additional reason for why such dynamics might lead to various suffering outcomes is the so-called Anna

Karenina principle (Diamond 1997, Zaneveld et al. 2017), named after the opening line of Tolstoy's novel *Anna Karenina:* "all happy families are all alike; each unhappy family is unhappy in its own way". The general form of the principle is that for a range of endeavors or processes, from animal domestication (Diamond 1997) to the stability of animal microbiomes (Zaneveld et al. 2017), there are many different factors that all need to go right, with even a single mismatch being liable to cause failure.

Within the domain of psychology, Baumeister et al. (2001) review a range of research areas to argue that "bad is stronger than good": while sufficiently many good events can overcome the effects of bad experiences, bad experiences have a bigger effect on the mind than good ones do. The effect of positive changes to well-being also tends to decline faster than the impact of negative changes: on average, people's well-being suffers and never fully recovers from events such as disability, widowhood, and divorce, whereas the improved well-being that results from events such as marriage or a job change dissipates almost completely given enough time (Lyubomirsky 2010).

To recap, various evolutionary and game-theoretical forces may push civilization in directions that are effectively random, random changes are likely to bad for the things that humans value, and the effects of bad events are likely to linger disproportionately on the human psyche. Putting these considerations together suggests (though does not guarantee) that freewheeling development could eventually come to produce massive amounts of suffering.

A possible counter-argument is that people are often more happy than their conditions might suggest. For example, as a widely-reported finding, while the life satisfaction reported by people living in bad conditions in slums is lower than that of people living in more affluent conditions, it is still higher than one might intuitively expect, and the slum-dwellers report being satisfied with many aspects of their life (Biswas-Diener & Diener 2001). In part, this is explained by fact that despite the poor conditions, people living in the slums still report many things that bring them pleasure: a mother who has lost two daughters reports getting joy from her surviving son, is glad that the son will soon receive a job at a bakery, and is glad about her marriage to her husband and feels that her daily prayer is important (Biswas-Diener & Diener 2001).

However, a proper evaluation of this research is complicated: "suffering" might be conceptualized as best corresponding to negative feelings, which are a separate component from cognitively evaluated life satisfaction (Lukas, Diener & Suh 1996), with the above slum-dweller study focusing mainly on life satisfaction. In general, life satisfaction is associated with material prosperity, while positive and negative feelings are associated with psychological needs such as autonomy, respect, and the ability to be able count on others in an emergency (Diener et al. 2010). A proper review of the literature and an analysis of how to interpret the research in terms of suffering risks lie beyond the scope of this paper.

4.3 Suffering outcome: cruelty-enabling technologies.

Better technology may enable people to better engage in cruel and actively sadistic pursuits. While active sadism and desire to hurt others may be a relatively rare occurrence in contemporary society, public cruelty has been a form of entertainment in many societies, ranging from the Roman practice of involuntary gladiator fights to animal cruelty in the Middle Ages. Even in contemporary society, there are widespread sentiments that people such as criminals should be severely punished in ways that inflict considerable suffering (part of the Roman gladiators were convicted criminals).

Contemporary society also contains various individuals who are motivated by the desire to hurt others (Torres 2016, 2017a, 2017b, ch. 4.), even to the point of sacrificing their own lives in the process. For example, Eric Harris, one of the two shooters of the Columbine High School Massacre, wrote extensively about his desire to rape and torture people, fantasized about tricking women into thinking that they were safe so that he could then hurt them, and wanted the freedom to be able to kill and rape without consequences (Langman 2015). While mass shooters tend to be lone individuals, there have existed more organized groups who seem to have given their members the liberty to act on similar motivations (Torres 2017a), such as the Aum Shinrikyo cult, where dissent or even just "impure thoughts" were punished by rituals amounting to torture and defectors "routinely kidnapped, tortured, imprisoned in cargo crates, subjected to electro shock, drugged in the Astral Hospital or killed outright" (Flannery 2016).

While most contemporary societies reject the idea of cruelty as entertainment, civilizations could eventually emerge in which such practices were again acceptable. Assuming advanced technology, this could take the form of keeping criminals and other undesirables alive indefinitely while subjecting them to eternal torture,[8] slaves kept for the purpose of sadistic actions who could be healed of any damage inflicted to them (one fictional illustration of such a scenario recently received

[8] Fictional depictions include Ellison (1967) and Ryding (no date); note that both stories contain very disturbing imagery. A third depiction was in the "White Christmas" episode of the TV series *Black Mirror,* which included a killer placed in solitary confinement for thousands of years while having to listen to a Christmas song on an endless loop.

widespread popularity as the TV series *Westworld*),[9] or even something like vast dystopian simulations of fantasy warfare inhabited by sentient "non-player characters", to serve as the location of massive multiplayer online games which people may play in as super-powered "heroes".

Particularly in the latter scenarios, the amount of sentient minds in such conditions could be many times larger than the civilization's other population. In contemporary computer games, it is normal for the player to kill thousands of computer-controlled opponents during the game, suggesting that a large-scale game in which a sizeable part of the population participated might instantiate very large numbers of non-player characters per player, existing only to be hurt for the pleasure of the players.

5 Suffering outcomes that may be caused by superintelligence[10]

In the previous section, we discussed possible suffering outcomes that might be realized without a singleton that could prevent them from occurring, and suggested that an appropriately-programmed superintelligence is currently the most likely candidate for forming such a singleton. However, an inappropriately programmed superintelligence could also cause suffering outcomes; we will now turn to this topic.

Superintelligence is related to three categories of suffering risk: *suffering subroutines* (Tomasik 2017), *mind crime* (Bostrom 2014) and *flawed realization* (Bostrom 2013).

5.1 Suffering subroutines

Humans have evolved to be capable of suffering, and while the question of which other animals are conscious or capable of suffering is controversial, pain analogues are present in a wide variety of animals. The U.S. National Research Council's Committee on Recognition and Alleviation of Pain in Laboratory Animals (2004) argues that, based on the state of existing evidence, at least all vertebrates should be considered capable of experiencing pain.

Pain seems to have evolved because it has a functional purpose in guiding behavior: evolution having found it suggests that pain might be the simplest solution for achieving its purpose. A superintelligence which was building subagents, such as worker robots or disembodied cognitive agents, might then also construct them in such a way that they were capable of feeling pain—and thus possibly suffering (Metzinger 2015)—if that was the most efficient way of making them behave in a way that achieved the superintelligence's goals.

Humans have also evolved to experience empathy towards each other, but the evolutionary reasons which cause humans to have empathy (Singer 1981) may not be relevant for a superintelligent singleton which had no game-theoretical reason to empathize with others. In such a case, a superintelligence which had no disincentive to create suffering but did have an incentive to create whatever furthered its goals, could create vast populations of agents which sometimes suffered while carrying out the superintelligence's goals. Because of the ruling superintelligence's indifference towards suffering, the amount of suffering experienced by this population could be vastly higher than it would be in e.g. an advanced human civilization, where humans had an interest in helping out their fellow humans.

Depending on the functional purpose of positive mental states such as happiness, the subagents might or might not be built to experience them. For example, Fredrickson (1998) suggests that positive and negative emotions have differing functions. Negative emotions bias an individual's thoughts and actions towards some relatively specific response that has been evolutionarily adaptive: fear causes an urge to escape, anger causes an urge to attack, disgust an urge to be rid of the disgusting thing, and so on. In contrast, positive emotions bias thought-action tendencies in a much less specific direction. For example, joy creates an urge to play and be playful, but "play" includes a very wide range of behaviors, including physical, social, intellectual, and artistic play. All of these behaviors have the effect of developing the individual's skills in whatever the domain. The overall effect of experiencing positive emotions is to build an individual's resources—be those resources physical, intellectual, or social.

To the extent that this hypothesis were true, a superintelligence might design its subagents in such a way that they had pre-determined response patterns for undesirable situations, so exhibited negative emotions. However, if it was constructing a kind of a command economy in which it desired to remain in control, it might not put a high value on any subagent accumulating individual resources. Intellectual resources would be valued to the extent that they contributed to the subagent doing its job, but physical and social resources could be irrelevant, if the subagents were provided with whatever resources necessary for doing their tasks. In such a case, the end result could be a world whose inhabitants experienced very little if any in the way of positive emotions, but did experience negative emotions. This could qualify as any one of the suffering outcomes we've considered (astronomical, net, pan-generational net).

One central and unresolved problem of suffering subroutines is the requirements for consciousness

[9] Another fictional depiction includes Gentle (2004); the warning for disturbing graphic imagery very much applies.

[10] This section reprints material that has previously appeared in a work by one of the authors (Gloor 2016), but has not been formally published before.

(Muehlhauser 2017) and suffering (Metzinger 2016, Tomasik 2017). The simpler the algorithms that can suffer, the more likely it is that an entity with no regard for minimizing it would happen to instantiate large numbers of them. If suffering has narrow requirements such as a specific kind of self-model (Metzinger 2016), then suffering subroutines may become less common.

Below are some pathways that could lead to the instantiation of large numbers of suffering subroutines (Gloor 2016):

Anthropocentrism. If the superintelligence were programmed to only care about humans, or by minds that were sufficiently human-like by some criteria, then it could end up being indifferent to the suffering of any other minds, including subroutines.

Indifference. If attempts to align the superintelligence with human values failed, it might not put any intrinsic value on avoiding suffering, so it may create large numbers of suffering subroutines.

Uncooperativeness. The superintelligence's goal is something like classical utilitarianism, with no additional regards for cooperating with other value systems. As previously discussed, classical utilitarianism would prefer to avoid suffering, all else being equal. However, this concern could be overridden by opportunity costs. For example, Bostrom (2003a) suggests that every *second* of delayed space colonization corresponds to a loss equal to 10^14 potential lives. A classical utilitarian superintelligence that took this estimate literally might choose to build colonization robots that used suffering subroutines, if this was the easiest way and developing alternative cognitive architectures capable of doing the job would take more time.

5.2 Mind crime

A superintelligence might run simulations of sentient beings for a variety of purposes. Bostrom (2014, p. 152) discusses the specific possibility of an AI creating simulations of human beings which were detailed enough to be conscious. These simulations could then be placed in a variety of situations in order to study things such as human psychology and sociology, and be destroyed afterwards.

The AI could also run simulations that modeled the evolutionary history of life on Earth in order to obtain various kinds of scientific information, or to help estimate the likely location of the "Great Filter" (Hanson 1998) and whether it should expect to encounter other intelligent civilizations. This could repeat the wild-animal suffering (Tomasik 2015, Dorado 2015) experienced in Earth's evolutionary history. The AI could also create and mistreat, or threaten to mistreat, various minds as a way to blackmail other agents.

As it is possible that minds in simulations could one day compose the majority of all existing minds (Bostrom 2003b)—and that, with sufficient technology, there could be astronomical numbers of them—then depending on the nature of the simulations and the net amount of happiness and suffering, mind crime could possibly lead to any one of the three suffering outcomes.

Below are some pathways that could lead to mind crime (Gloor 2016):

Anthropocentrism. Again, if the superintelligence were programmed to only care about humans, or about minds that were sufficiently human-like by some criteria, then it could be indifferent to the suffering experienced by non-humans in its simulations.

Indifference. If attempts to align the superintelligence with human values failed, it might not put any intrinsic value on avoiding suffering, and may thus create large numbers of simulations with sentient minds if that furthered its objectives.

Extortion. The superintelligence comes into conflict with another actor that disvalues suffering, so the superintelligence instantiates large numbers of suffering minds as a way of extorting the other entity.

Libertarianism regarding computations: the creators of the first superintelligence instruct the AI to give every human alive at the time control of a planet or galaxy, with no additional rules to govern what goes on within those territories. This would practically guarantee that some humans would use this opportunity for inflicting widespread cruelty (see the previous section).

5.3 Flawed realization

A superintelligence with human-aligned values might aim to convert the resources in its reach into clusters of utopia, and seek to colonize the universe in order to maximize the value of the world (Bostrom 2003a), filling the universe with new minds and valuable experiences and resources. At the same time, if the superintelligence had the wrong goals, this could result in a universe filled by vast amounts of *disvalue*.

While some mistakes in value loading may result in a superintelligence whose goal is completely unlike what people value, certain mistakes could result in *flawed realization* (Bostrom 2013). In this outcome, the superintelligence's goal gets human values *mostly* right, in the sense of sharing many similarities with what we value, but also contains a flaw that drastically changes the intended outcome.[11]

[11] One fictional illustration of a flawed utopia is Yudkowsky (2009), though this setting does not seem to contain enormous amounts of suffering.

For example, value-extrapolation (Yudkowsky 2004) and value-learning (Soares 2016, Sotala 2016) approaches attempt to learn human values in order to create a world that is in accordance with those values. There have been occasions in history when circumstances that cause suffering have been defended by appealing to values which seem pointless to modern sensibilities, but which were nonetheless a part of the prevailing values at the time. In Victorian London, the use of anesthesia in childbirth was opposed on the grounds that being under the partial influence of anesthetics may cause "improper" and "lascivious" sexual dreams (Farr 1980), with this being considered more important to avoid than the pain of childbirth.

A flawed value-loading process might give disproportionate weight to historical, existing, or incorrectly extrapolated future values whose realization then becomes more important than the avoidance of suffering. Besides merely considering the avoidance of suffering less important than the enabling of other values, a flawed process might also tap into various human tendencies for endorsing or celebrating cruelty (see the discussion in section 4), or outright glorifying suffering. Small changes to a recipe for utopia may lead to a future with much more suffering than one shaped by a superintelligence whose goals were completely different from ours.

6 How and whether to work on s-risk?

In the previous sections, we have argued for s-risks being severe enough to be worth preventing, and for the existence of several plausible routes by which they might be realized. We will now argue for the case that it is possible to productively work on them today, via some of the following recommendations.

Carry out general AI alignment work. Given that it would generally be against the values of most humans for suffering outcomes to be realized, research aimed at aligning AIs with human values (Yudkowsky 2008, Goertzel & Pitt 2012, Bostrom 2014, Sotala 2016, Soares & Fallenstein 2017) seems likely to also reduce the risk of suffering outcomes. If our argument for suffering outcomes being something to avoid is correct, then an aligned superintelligence should also attempt to establish a singleton that would prevent negative suffering outcomes, as well as avoiding the creation of suffering subroutines and mind crime.

In addition to technical approaches to AI alignment, the possibility of suffering risks also tends to make more similar recommendations regarding social and political approaches. For example, Bostrom et al. (2016) note that conditions of *global turbulence* might cause challenges for creating value-aligned AI, such as if pre-existing agreement are not kept to and ill-conceived regulation is enacted in a haste. Previous work has also pointed to the danger of arms races making it harder to keep AI aligned (Shulman 2009, Miller 2012, Armstrong et al. 2013). As the avoidance of suffering outcomes is the joint interest of many different value systems, measures that reduce the risk of arms races and improve the ability of different value systems to shape the world in their desired direction can also help avoid suffering outcomes.

Besides making AIs more aligned in general, some interventions may help avoid negative outcomes—such as suffering outcomes from flawed realization scenarios—in particular. Most of the current alignment research seeks to ensure that the values of any created AIs are aligned with humanity's values to a maximum possible extent, so that the future they create will contain as much positive value as possible. This is a difficult goal: to the extent that humanity's values are complex and fragile (Yudkowsky 2011), successful alignment may require getting a very large amount of details right.

On the other hand, it seems much easier to give AIs goals that merely ensure that they will not create a future with *negative* value by causing suffering outcomes. This suggests an approach of fail-safe methods: safety nets or mechanisms such that, if AI control fails, the outcome will be as good as it gets under the circumstances. Fail-safe methods could include tasking AI with the objective of buying more time to carefully solve goal alignment more generally, or fallback goal functions:

Research fallback goals: Research ways to implement multi-layered goal functions, with a "fallback goal" that kicks in if the implementation of the top layer does not fulfill certain safety criteria. The fallback would be a simpler, less ambitious goal that is less likely to result in bad outcomes. Difficulties would lie in selecting the safety criteria in ways that people with different values could all agree on, and in making sure that the fallback goal gets triggered under the correct circumstances.

Care needs to be taken with the selection of the fallback goal, however. If the goal was something like reducing suffering, then in a multipolar (Bostrom 2014) scenario, other superintelligences could have an incentive to create large amounts of suffering in order to coerce the superintelligence with the fallback goal to act in some desired way.

Research ways to clearly separate superintelligence designs from ones that would contribute to suffering risk. Yudkowsky (2017) proposes building potential superintelligences in such a way as to make them widely separated in design space from ones that would cause suffering outcomes. For example, if an AI has a representation of "what humans value" *V* which it is trying to maximize, then it would only take a small (perhaps accidental) change to turn it into one that maximized *-V* instead, possibly causing enormous suffering. One proposed way of achieving this is by never trying to explicitly represent complete human values: then, the AI "just doesn't contain the information needed to compute states of the universe that we'd consider worse than death; flipping the sign of the utility function *U*, or subtracting components from *U* and then flipping the sign, doesn't identify any state we consider worse than [death]" (Yudkowsky 2017). This would also reduce the risk of suffering being created through another actor that was trying to extort the superintelligence.

Carry out research on suffering risks and the enabling factors of suffering. At this moment, there is only little research to the possibility of risks of astronomical suffering. Two kinds of research would be particularly useful. First, research focused on understanding the biological and algorithmic foundation of suffering (Metzinger 2016) could help understand how likely outcomes such as suffering subroutines would be. Pearce (1995) has argued for the possibility of minds motivated by "gradients of bliss", which would not need to experience any suffering: if minds could be designed in such a manner, it might help avoid suffering outcomes.

Second, research on suffering outcomes in general, to understand how to avoid them. With regard to suffering risks from extortion scenarios, targeted research in economics, game theory or decision theory could be particularly valuable.

Rethink maxipok and maximin. Bostrom (2002, 2013) proposes a "maxipok rule" to act as a rule of thumb when trying to act in the best interest of humanity as a whole:

Maxipok: Maximize the probability of an "OK outcome", where an OK outcome is any outcome that avoids existential catastrophe.

The considerations in this paper do not necessarily refute the rule as written, especially not since Bostrom defines an "existential catastrophe" to include "the permanent and drastic destruction of its potential for desirable future development", and the realization of suffering outcomes could very well be thought to fall under this definition. However, in practice much of the discourse around the concept of existential risk has focused on the possibility of extinction, so it seems valuable to highlight the fact that "existential catastrophe" does not include only scenarios of zero value, but also scenarios of negative value.

Bostrom (2002, 2013) also briefly discusses the "maximin" principle, "choose the action that has the best worst-case outcome", and rejects this principle as he argues that this entails "choosing the action that has the greatest benefit under the assumption of impending extinction. Maximin thus implies that we ought all to start partying as if there were no tomorrow." (Bostrom 2013, p. 19). However, since a significant contribution to the expected value of AI comes from worse outcomes than extinction, this argument is incorrect. While there may be other reasons to reject maximin, the principle correctly implies choosing the kinds of actions that avoid the worst suffering outcomes and so might not be very dissimilar from maxipok.

7 Acknowledgments

David Althaus, Stuart Armstrong, Tobias Baumann, Max Daniel, Ruairi Donnelly, and two anonymous reviewers provided valuable feedback on this paper.

8 References

Alexander, S. (2014) Meditations on Moloch. *Slate Star Codex.* http://slatestarcodex.com/2014/07/30/meditations-on-moloch/

Armstrong, S., & Sandberg, A. (2013). Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica, 89,* 1-13.

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. *AI & Society, 31*(2), 201-206.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370.

Benatar, D. (2008). *Better never to have been: the harm of coming into existence.* Oxford University Press.

Biswas-Diener, R. & Diener, E. (2001). Making the best of a bad situation: Satisfaction in the slums of Calcutta. *Social Indicators Research*, 55, 329-352.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9*(1).

Bostrom, N. (2003a). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas, 15*(3), 308-314.

Bostrom, N. (2003b). Are we living in a computer simulation? *The Philosophical Quarterly*, 53(211), 243-255.

Bostrom, N. (2004). The future of human evolution. In Tandy, C. (ed.) *Death and anti-death: Two hundred years after Kant, fifty years after Turing*, 339-371.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy, 4*(1), 15-31.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* OUP Oxford.

Bostrom, N., Dafoe, A., & Flynn, C. (2016) Policy Desiderata in the Development of Machine

Superintelligence. https://nickbostrom.com/papers/aipolicy.pdf

Diamond, J. (1997). *Guns, Germs, and Steel: The Fates of Human Societies.* W. W. Norton.

Diener, E., Ng, W., Harter, J. & Arora, R. (2010). Wealth and Happiness Across the World: Material Prosperity Predicts Life Evaluation, Whereas Psychosocial Prosperity Predicts Positive Feeling. *Journal of Personality and Social Psychology, 99(1)*, 52– 61.

Ditto, P. H., Druley, J. A., Moore, K. A., Danks, J. H., & Smucker, W. D. (1996). Fates worse than death: the role of valued life activities in health-state evaluations. *Health Psychology, 15*(5), 332.

Dorado, D. (2015). Ethical Interventions in the Wild: An Annotated Bibliography. *Relations: Beyond Anthropocentrism, 3*, 219.

Ellison, H. (1967). I Have No Mouth, and I Must Scream. *IF: Worlds of Science Fiction*, March 1967.

Farr, A. D. (1980). Early opposition to obstetric anaesthesia. *Anaesthesia, 35*(9), 896-907.

Fehige, C. (1998) “A Pareto Principle for Possible People” in Fehige, C. & Wessels, U. (eds.) *Preferences.* Berlin: De Gruyter, 509-543.

Food and Agriculture Organization of the United Nations. (2012). FAOSTAT Agriculture – Livestock Primary Dataset: world total of animals slaughtered for meat in 2012. Retrieved January 26, 2016, from http://faostat.fao.org/site/569/DesktopDefault.aspx?PageID=569

Flannery, F. (2016). Understanding Apocalyptic Terrorism: Countering the Radical Mindset. Abingdon, Oxon: Routledge.

Fredrickson, B. L. (1998). What good are positive emotions? *Review of General Psychology, 2*(3), 300.

Gentle, M. (2004) Human Waste. In *Cartomancy*. Gollancz.

Gloor, L. (2016). Suffering-focused AI safety. Foundational Research Institute. https://foundational-research.org/suffering-focused-ai-safety-why-fail-safe-measures-might-be-our-top-intervention/

Gloor, L. (2017). Tranquilism. Foundational Research Institute. https://foundational-research.org/tranquilism/

Gloor, L. & Mannino, A. (2016). The Case for Suffering-Focused Ethics. Foundational Research Institute. https://foundational-research.org/the-case-for-suffering-focused-ethics/

Goertzel, B. & Pitt, J., (2012). Nine ways to bias open-source AGI toward friendliness. *Journal of Evolution and Technology, 22*(1).

Greaves, H (2017). Population axiology. *Philosophy Compass*, 12:e12442. https://doi.org/10.1111/phc3.12442

Greene, J. (2013). *Moral tribes: Emotion, Reason, and the gap between us and them.* Penguin.

Hanson, R. (1998). The Great Filter - Are We Almost Past It? http://mason.gmu.edu/~rhanson/greatfilter.html

Hanson, R. (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth.* Oxford University Press.

Häggström, O. (2016). *Here Be Dragons: Science, Technology and the Future of Humanity.* Oxford University Press.

Langman, P. (2015) *School Shooters: Understanding High School, College, and Adult Perpetrators*. Lanham, MD: Rowman & Littlefield.

Le Guin, U. K. (1973). The ones who walk away from Omelas. In Silverberg, R. (ed.), *New Dimensions 3.*

Lukas, R.E., Diener, E. & Suh, E. (1996). Discriminant Validity of Well-Being Measures. *Journal of Personality and Social Psychology*, 71(3), 616-628.

Lyubomirsky, S. (2010). 11 Hedonic Adaptation to Positive and Negative Experiences. In Folkman, S. & Nathan, P.E. (eds.) *The Oxford Handbook of Stress, Health, and Coping*. Oxford University Press.

Manning, P. (1992) The Slave Trade: The Formal Demography of a Global System. In Klein, M. A., & Hogendorn, J. (eds.) *The Atlantic slave trade: effects on economies, societies and peoples in Africa, the Americas, and Europe*. Duke University Press.

Metzinger, T. (2015). What if they need to suffer? Edge.org response to: What do you think about machines that think? https://www.edge.org/response-detail/26091

Metzinger, T. (2016) Suffering. In Almqvist, K. & Haag, A. (eds.) *The Return of Consciousness*. Stockholm: Axel and Margaret Ax:son Johnson Foundation

Miller, J. D. (2012). *Singularity Rising: Surviving and thriving in a smarter, richer, and more dangerous world*. BenBella Books, Inc.

Muehlhauser, L. (2017). 2017 Report on Consciousness and Moral Patienthood. *Open Philanthropy Project.* https://www.openphilanthropy.org/2017-report-consciousness-and-moral-patienthood

Narveson, J. (1973). Moral problems of population. *The Monist*, 62-86.

National Research Council's Committee on Recognition and Alleviation of Pain in Laboratory Animals. (2009). Recognition and alleviation of pain in laboratory animals. Washington (DC): National Academies Press .

Nussbaum, M. (1997) Capabilities and Human Rights, 66 Fordham L. Rev. 273.

Ord, T. (2013). Why I'm Not a Negative Utilitarian. http://www.amirrorclear.net/academic/ideas/negative-utilitarianism/

Parfit, D. (1991). Equality or Priority? (Department of Philosophy: University of Kansas)

Pearce, D. (1995) The Hedonistic Imperative. https://www.hedweb.com/hedab.htm

Persson, I., & Savulescu, J. (2012). *Unfit for the future: the need for moral enhancement.* Oxford University Press.

Rachels, S (1998). Counterexamples to the transitivity of better than. *Australasian Journal of Philosophy,* 76(1):71-83.

Regan, T (1980). Utilitarianism, Vegetarianism, and Animal Rights. *Philosophy and Public Affairs*, 9(4):305-324.

Rubin, E. B., Buehler, A. E., & Halpern, S. D. (2016). States worse than death among hospitalized patients with serious illnesses. *JAMA Internal Medicine, 176*(10), 1557-1559.

Ryding, D. (no date). Yes, Jolonah, There Is A Hell. *Orion's Arm.* http://www.orionsarm.com/page/233

Sayah, F. A., Mladenovic, A., Gaebel, K., Xie, F., & Johnson, J. A. (2015). How dead is dead? Qualitative findings from participants of combined traditional and lead-time time trade-off valuations. *Quality of Life Research, 25*(1), 35-43.

Shulman, C. (2012) Are pain and pleasure equally energy-efficient? **Reflective Disequilibrium**. https://reflectivedisequilibrium.blogspot.fi/2012/03/are-pain-and-pleasure-equally-energy.html

Shulman, C., & Armstrong, S. (2009). Arms control and intelligence explosions. In *7th European Conference on Computing and Philosophy (ECAP)*, Bellaterra, Spain, July (pp. 2-4).

Singer, P. (1981/2011). *The Expanding Circle: Ethics, Evolution, and Moral Progress.* Princeton University Press.

Singer, P. (1993). *Practical Ethics*, second edition. Cambridge University Press.

Soares, N., & Fallenstein, B. (2017). Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In Callaghan et al. (eds.) *The Technological Singularity - Managing the Journey* (pp. 103-125). Springer Berlin Heidelberg.

Soares, N. (2016). The Value Learning Problem. *2nd International Workshop on AI and Ethics, AAAI-2016.* Phoenix, Arizona.

Sotala, K., & Yampolskiy, R. V. (2015). Responses to Catastrophic AGI Risk: A Survey. *Physica Scripta, 90*(1), 018001.

Sotala, K. (2016). Defining Human Values for Value Learners. *2nd International Workshop on AI and Ethics, AAAI-2016.* Phoenix, Arizona.

Tomasik, B. (2015). The importance of wild-animal suffering. *Relations: Beyond Anthropocentrism, 3*, 133.

Tomasik, B. (2017) What Are Suffering Subroutines? http://reducing-suffering.org/what-are-suffering-subroutines/

Torres, P. (2016) Agential Risks: A Comprehensive Introduction. Journal of Evolution and Technology, 26(2), pp. 31-47.

Torres, P. (2017a) Who Would Destroy the World? Omnicidal Agents and Related Phenomena. Pre-publication draft: https://docs.wixstatic.com/ugd/d9aaad_b18ce62c32be44ddbf64268fc295fdc0.pdf

Torres, P. (2017b) *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks.* Durham, North Carolina: Pitchstone Publishing.

West, B. (2017) An Argument for Why the Future May Be Good. *Effective Altruism Forum*, http://effective-altruism.com/ea/1cl/an_argument_for_why_the_future_may_be_good/ .

World Food Programme. (2017). Zero Hunger. http://www1.wfp.org/zero-hunger

Yudkowsky, E. (2004). Coherent extrapolated volition. Singularity Institute for Artificial Intelligence. https://intelligence.org/files/CEV.pdf

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In Bostrom, N. & Ćirković, M.M. (eds.) *Global Catastrophic Risks.* New York: Oxford University Press.

Yudkowsky, E. (2009) Failed Utopia #4-2. *Less Wrong.* http://lesswrong.com/lw/xu/failed_utopia_42/

Yudkowsky, E. (2011). Complex Value Systems are Required to Realize Valuable Futures. Machine Intelligence Research Institute. https://intelligence.org/files/ComplexValues.pdf

Yudkowsky, E. (2017) Separation from hyperexistential risk. *Arbital.* Retrieved December 11, 2017, from https://arbital.com/p/hyperexistential_separation/

Zaneveld, J. R., McMinds, R., & Vega, T. R. (2017). Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nature Microbiology, 2*, 17121.