AVOIDING THE WORST

How to Prevent a Moral Catastrophe

TOBIAS BAUMANN

Parts of this book have previously been published elsewhere (in modified form) by the author.

Contents

Introduction

Human history is full of moral catastrophes: centuries of slavery, devastating wars, oppressive tyrants, cruel genocides. The range of atrocities fills many books.[1] In many cases, the actions and moral beliefs of the past were horrifying by today's standards.

We like to think that we have put all that behind us, but we have not. While there has undoubtedly been some progress, wars and slavery are, on a global perspective, definitely not things of the past.[2] And just as past generations often failed to realise how wrong their actions were, we may fail to recognise a contemporary moral catastrophe due to the moral blind spots of our time.[3] For instance, we raise and kill vast numbers of animals each year on factory farms and in slaughterhouses, often inflicting terrible suffering on them in the process. If we take the argument for animal rights seriously, as I believe we should, then this constitutes an ongoing moral catastrophe.[4]

Other writers have documented and analysed these issues in depth. But what about the possibility of a *future* moral catastrophe? Could such tragedies potentially take place on an even larger scale? And what can we do now to prevent that from happening? These questions have not yet been explored in much depth, and my book

---

[1] White, 2011 provides an overview of the 100 worst atrocities.

[2] There is an ongoing academic debate on historical trends in the frequency and intensity of wars. Pinker, 2011 argues that war has declined substantially (along with other forms of violence), while Braumoeller, 2019 takes the opposite view.
Similarly, it is sometimes claimed that there are now more slaves than ever before. However, this is highly uncertain and depends on how slavery is defined and measured. Cf. Gerrard, 2020.

[3] Cf. Williams, 2015.

[4] See Horta, 2022 for more details on the many forms of suffering that animals have to endure, as well as a thorough argument for why nonhuman beings matter morally.

aims to fill this gap.

I approach this topic with the belief that we should use our limited resources to help others as effectively as possible.[5] From this perspective, some of the most important questions concern the scale and the likelihood of future moral catastrophes. If human civilisation will have advanced technology at its disposal without sufficient *moral* progress to use it responsibly, we risk causing unprecedented levels of suffering. Likewise, an expansion into space could increase the amount of suffering by many orders of magnitude. This astronomical scope is a strong reason to take the risk of worst-case outcomes seriously, even if the likelihood remains unclear. I revisit these themes throughout the book.

Many readers may feel a tension between such abstract thinking about future risks and the urgent desire to do something to prevent horrible suffering in the here and now. I feel this tension myself, and the drive to help immediately is laudable. At the same time, our drive to help should not prevent us from thinking critically about what is most impactful in the big picture. At least, we should keep an open mind and explore the risk of a future moral catastrophe.

Some readers might likewise find it unpleasant to think in depth about worst-case futures. It can be disturbing to think about scenarios that involve a lot of suffering, yet we cannot afford to ignore the risk of such catastrophic scenarios if we want to do as much good as possible. We must objectively consider the arguments and the available information, however worrisome they may be.[6]

Before I dive deeper, I should clarify the values that underlie this book. A key principle is *impartiality*: suffering matters equally irrespective of *who* experiences it. In particular, I believe we should care about all sentient beings, including nonhuman animals.[7] Similarly, I believe suffering matters equally regardless of *when* it is experienced. A future individual is no less (and no more) deserving of

---

[5] This idea has been termed *effective altruism*. See MacAskill, 2015; Vinding, 2018d.

[6] Of course, it is also quite possible that the future will be much better than today, and it could be (almost) free of suffering if things go well. However, such optimistic scenarios are not the focus of this book.

[7] See Singer, 1975; Horta, 2010; Vinding, 2015; Horta, 2022.

moral consideration than someone alive now. So the fact that a moral catastrophe takes place in the distant future does not reduce the urgency of preventing it, if we have the means to do so.[8] I will assume that you broadly agree with these fundamental values, which form the starting point of the book.

The book is divided into three parts. Part 1 lays the conceptual groundwork by introducing the notion of risks of astronomical suffering (*s-risks*), which forms the centrepiece of the book. I provide a definition to distinguish s-risks from other bad future outcomes, outline different types of s-risks, and give examples of how s-risks could come about.

In Part 2 of the book, I review arguments for and against prioritising the reduction of s-risks. I break this question down into three subquestions: whether we should focus on the long-term future, whether we should focus on averting suffering, and whether we should focus on preventing worst-case outcomes. In addition, I discuss potential biases that might distort our thinking on these questions.

In Part 3, I explore how we can best reduce s-risks. I outline plausible interventions along with their advantages and their drawbacks. Finally, I conclude with a discussion of how to proceed in light of great uncertainty about the future.

[8] See MacAskill, 2022, Section "Future People Count" for an argument in favour of taking future individuals into account.

Part I

What are s-risks?

CHAPTER ONE

Technology and astronomical stakes

Throughout human history, the emergence of new technologies has often had a transformative impact on society. We reap the fruits of technological progress every day. Our smartphones would seem like magic to people living just a century ago. But perhaps more importantly, we live longer than ever before, we have managed to eradicate many diseases, and we are, at least on average, vastly richer than past generations.

Yet there is another side to the story. Technology also brought with it industrial warfare,[9] the atomic bomb, and environmental disasters. While new technologies offer unprecedented opportunities, they also pose serious risks, especially when combined with insufficient moral progress.

The risks are exacerbated when we consider nonhuman animals. Industrialisation has increased the consumption of meat and other animal products. This has multiplied the number of animals who are raised and killed, usually in deplorable conditions on factory farms

[9] This is not to say that industrial warfare is necessarily worse than earlier forms of warfare. In fact, it has been argued that the number of military casualties (per capita) has decreased over time. See Pinker, 2011.

and in industrial slaughterhouses.[10] And it is worth noting that this is not due to intentional malice — after all, most people do not approve of animal suffering.[11] Instead, factory farming is mainly the result of economic incentives and technological feasibility, coupled with a lack of moral concern.

Barring extinction or civilizational collapse, technological progress will likely continue and endow humanity with new capabilities. If such advances allow us to expand into space and colonise other planets, the stakes will become truly *astronomical.* It is conceivable that human civilisation could eventually populate billions of galaxies.[12]

If Earth-originating civilisation develops advanced technology or spreads out into space, it is more important than ever that we use our new capabilities responsibly. We need to be mindful of the possibility that future technologies might, coupled with indifference, lead to a moral catastrophe of colossal proportions. With so much at stake, we must close the gap between our power and our wisdom.

The concept of s-risks

In this book, I consider the risk of a future that contains vast quantities of suffering. Such scenarios have been labelled ***risks of astronomical suffering***[13], but for brevity I will mostly use the short form ***suffering risks*** or ***s-risks***.

Formally, s-risks have been defined as "risks of events that bring about suffering in cosmically significant amounts", where "significant" means "significant relative to expected future suffering".[14] In less formal terms, s-risks are scenarios that involve severe suffering on an

---

[10] See Horta, 2022, Chapter 3 for more details on how animals are harmed (both in animal agriculture and in other areas), and Ritchie & Roser, 2017 for an overview of the number of animals slaughtered per year.

[11] Sentience Institute, 2017.

[12] According to estimates, 100-400 billion stars exist in our galaxy (the Milky Way) alone, and there are 100-200 billion galaxies throughout the universe. Cf. Howell, 2021; Howell & Harvey, 2022. For more on the question of whether colonisation is feasible, see Beckstead, 2014.

[13] One of the first documented uses of the term is Tomasik, 2011.

[14] Althaus & Gloor, 2016.

astronomical scale, vastly exceeding all suffering that has existed on Earth so far. You can imagine a future development akin to factory farming, but on an even more horrendous scale.

Note that the concept of s-risks is gradual rather than binary. That is, s-risks can be more or less severe. Also worth noting is that the definition of s-risks refers to *absolute* amounts of suffering rather than to the ratio of suffering to overall population size. So a scenario in which the population size is extremely large can count as an s-risk even if only a small fraction of the population is affected, as long as the total amount of suffering is sufficiently high.

S-risks, dystopia, and x-risks

The concept of s-risks is more specific than "any scenario that is considered (very) bad". For instance, climate change is not an s-risk. While climate change causes a well-documented range of adverse effects, from wildfires to sea level rise,[15] it need not result in astronomical quantities of suffering *per se*.[16]

Similarly, s-risks should be distinguished from the related but less specific notion of *dystopia.* Both terms are about worst-case outcomes, but "dystopia" is a broad term that can refer to any (hypothetical) future society that is considered highly undesirable. Since vast quantities of suffering are surely highly undesirable, s-risks can be viewed as a class of dystopian scenarios.[17]

However, not every dystopian scenario qualifies as an s-risk. This is primarily because the definition of s-risks involves an astronomical scale, whereas a dystopia might take place on a smaller scale. In

---

[15] Intergovernmental Panel on Climate Change, 2022.

[16] One could perhaps argue that climate change might still indirectly precipitate s-risks, such as by exacerbating conflicts. And even if climate change is not an s-risk, that does not mean that we should do nothing about it, since s-risks are only one class of risks among many.

[17] This categorisation can depend on one's values. For those who think that vast amounts of suffering can be outweighed by happiness or other (purported) goods, some s-risk scenarios may not be considered dystopian (if they also contain a sufficient quantity of positive goods). I discuss this in more detail in Chapter 4.

addition, the concept of s-risks focuses on sheer suffering, whereas many commonly discussed dystopian scenarios (e.g. George Orwell's *Nineteen Eighty-Four*) emphasise themes such as a tyrannical government, surveillance, or a loss of freedom. One may consider something a dystopia, but not an s-risk, even if the population does not suffer severely — e.g., because of brainwashing or ubiquitous entertainment. This highlights the subjectiveness of what one considers dystopian, which is part of why I prefer the less ambiguous concept of s-risk.

To avoid confusion, we also need to distinguish s-risks from the concept of *existential risk*. Existential risks, often abbreviated as x-risks, are defined as "an adverse outcome that would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential".[18]

How do s-risks relate to x-risks? Both are about the long-term future of civilisation and the possibility of affecting outcomes on an astronomical scale. But x-risks are about humanity's potential, while s-risks are about outcomes that contain vast amounts of suffering.

There is some overlap between the two concepts. For instance, if a totalitarian dictator takes over the world, this might both curtail humanity's potential (x-risk) and cause vast amounts of suffering (s-risk).

Still, s-risks are a distinct concept and not a subclass of x-risks. It is conceivable to have astronomical amounts of suffering that neither lead to extinction nor curtail humanity's "potential".[19] Some forms of suffering, especially ones that affect nonhuman beings, may even be economically useful to human civilization. So not all s-risks are x-risks.[20]

---

[18] Bostrom, 2002.

[19] To be precise, this depends on one's interpretation of the rather vague term "potential". For example, if "potential" refers to the potential of a utopian future without any suffering, then every s-risk is (by definition) an x-risk, too. But if "potential" refers to, say, a future that contains a vast number of humans, of whom most are satisfied with their lives, then the realization of humanity's potential does not rule out s-risks.

[20] Neither does the converse hold. For example, it would be an x-risk, but not an s-risk, if human civilisation goes extinct and the universe remains empty. See also Aird, 2020. (This brackets complicated questions about whether other

S-risks, artificial intelligence, and artificial sentience

I have mentioned the idea that new technologies can have a massive impact on society. But what are specific candidates of transformative future technologies? One common hypothesis is that advanced artificial intelligence (AI) is likely to be a crucial technology.[21] If we create systems that can surpass humans in terms of general intelligence, then this would (so the argument goes) constitute a pivotal event in human history.

In particular, advanced AI systems could cause s-risks if their increasing capabilities are not managed in a prudent way — and conversely, they hold the potential to prevent s-risks from other sources.[22] If advanced AI will be a pivotal technology, we should perhaps analyse s-risks primarily in the context of scenarios in which advanced AI will determine the future. This raises many complex questions, which I will discuss in more detail later (in Chapter 10).

Of course, the future of artificial intelligence is highly uncertain, and it is unclear how we can positively influence such scenarios. Given this uncertainty, I will not focus exclusively on s-risks caused by advanced AI. But we should keep the possibility in mind that increasingly powerful AI systems could play a crucial role in shaping the long-term future.

A second way in which advanced AI might be relevant relates to artificial *sentience* rather than intelligence. Artificial entities may, so the argument goes, develop a capacity for subjective experience at some point. The range of conceivable forms of artificial sentience includes digital enhancements to human bodies, "uploading" of minds into so-called "whole brain emulations", or sentient simulations taking place within a larger computer program.

Consider this thought experiment:[23] If you were to take a sentient biological brain, and replace one neuron after another with a

---

civilisations exist or will come into existence in the future.)

[21] Bostrom, 2014.

[22] Sotala & Gloor, 2017.

[23] This is often called the "fading qualia argument". See Chalmers, 1995.

functionally equivalent computer chip, would it make the brain less sentient? Would the brain still be sentient once all of its biological neurons have been replaced? If not, at what point would it cease to be sentient? This raises various philosophical questions about the nature of consciousness. A detailed review is beyond the scope of this book, but I would note that most theories tend to agree that artificial sentience is possible in principle.[24]

To be clear, we do not know whether sentient artificial minds will actually be instantiated. Perhaps technology will never evolve in that direction, even if artificial sentience is possible in principle. But *if* conscious artificial beings come into existence, their well-being matters morally.[25] It would not be justifiable to disregard their interests (and potential suffering) merely because such beings are based on silicon rather than carbon.

At the same time, many forms of artificial beings will likely be very alien to us and therefore pose a unique challenge to our moral sentiments. Such disembodied minds may have no face, body movements, or screams to which we can relate. That makes it difficult to empathize with them on a visceral level. In addition, we do not yet have a reliable way to detect sentience, especially in systems that are fundamentally different from human brains. We might therefore fail to recognise sentience and suffering in such "voiceless" beings.[26]

For all these reasons, it seems likely that people will not care to a sufficient extent about the well-being of (some forms of) artificial minds, provided that such minds are created. The human record of cruelty towards other humans and towards nonhuman animals does not bode well for future artificial minds.

This is particularly worrisome if large numbers of sentient artificial minds are created in the future. One reason this might happen is that artificial minds are likely to have significant advantages over biological

---

[24] See Muehlhauser, 2017 for more details.

[25] For an overview of existing work on the moral consideration of artificial entities, see Harris & Anthis, 2021.

[26] It is worth noting that many philosophers and scientists also failed to recognise animal sentience for thousands of years, despite the much greater degree of similarity between nonhuman animals and humans.

minds, making them economically expedient.[27] Also, just as one can mass produce hardware and copy computer programs at will, it is plausible that creating large numbers of artificial minds will be very easy.

The combination of potentially vast numbers of sentient artificial minds and the foreseeable lack of empathy poses a serious s-risk. In fact, these conditions look strikingly similar to those of factory farming, which is also characterised by economic expediency in combination with moral indifference towards nonhuman beings.

[27] Sotala, 2012.

CHAPTER TWO

Types of s-risks

In this chapter, I turn to the question of how s-risks could come about. This is a challenging question for many reasons. In general, it is hard to imagine what developments in the distant future might look like. Scholars in the Middle Ages could hardly have anticipated the atomic bomb. Given this great uncertainty, the following examples of s-risks are to be understood mostly as informed guesses for illustrative purposes, and not as a claim that these scenarios constitute the most likely s-risks. All of these examples put together might still account for only a small fraction of s-risks compared to currently *unknown* s-risks.

Going into detail on any specific scenario also carries a risk of *availability bias*. The availability bias is the tendency to view an example that comes readily to mind as more representative than it really is.[28] This mental shortcut can be problematic if, as seems plausible, s-risks are actually distributed over a broad range of possible sources and scenarios. To counteract the availability bias, I will briefly outline many possible scenarios, rather than going into detail on any specific one.

Yet another challenge is the lack of clear feedback loops, especially if the scenarios in question have no comparable precedent in history. A conventional evidence-based approach is therefore not straightforwardly applicable.

Given these challenges, it is most fruitful to explore many different types of s-risks. I will group the spectrum of possible scenarios into

[28] Cf. Tversky & Kahneman, 1973.

three categories: **incidental s-risks**, **agential s-risks**, and **natural s-risks**.

Incidental s-risks

Incidental s-risks arise when efficient ways to achieve a certain goal creates a lot of suffering in the process, without anyone actively trying to cause suffering per se. The agent or agents that cause the s-risk are either indifferent to that suffering, or they would prefer a suffering-free alternative in theory, but are not willing to bear the necessary costs in practice.

We can further divide incidental s-risks into subcategories based on the underlying motivation. In one class of scenarios, economic incentives and market forces cause large amounts of suffering because that suffering is a byproduct of high economic productivity. We have already encountered this in the case of factory farming. It just so happens that the most economically efficient way to satisfy the demand for cheap animal products entails miserable conditions for farmed animals.[29] Future technology might enable similar dynamics but on a much larger scale.

Another possibility involves suffering that is instrumental for information gain. Experiments on sentient creatures can be useful for scientific purposes, while causing serious harm to those experimented on. Again, future technology may enable such practices on a much larger scale. As discussed in the previous chapter, it may become possible to run a large number of simulations of artificial minds that are capable of suffering. And if there are instrumental reasons to run many such simulations, this could lead to vast amounts of suffering. For example, an advanced AI system might run many simulations to improve its knowledge of human psychology or in an attempt to predict what other agents will do in a certain situation.

It is also conceivable that complex simulations will be used for entertainment purposes in the future, which could cause serious

[29] Animal agriculture does not qualify as an example of a (realised) s-risk because an s-risk requires astronomical scope (by definition). That is, the amount of suffering would need to exceed current suffering by several orders of magnitude for something to constitute an s-risk.

suffering if these simulations contain artificially sentient beings. Many people enjoy violent forms of entertainment, as evidenced by countless historical examples, from gladiator fights to public executions. Another case in point is the content of today's video games or movies. Such forms of entertainment are victimless as long as they are fictional — but in combination with sentient artificial minds, they could be a potential s-risk.

Agential s-risks

The previous examples are situations in which an efficient solution to a problem happens to involve a lot of suffering as an unintentional byproduct. A different class of s-risks, which I call **agential s-risks**, arises when an agent actively and intentionally wants to cause harm.

A simple example is sadism. A minority of future agents might, like some humans, derive pleasure from inflicting pain on others. This will hopefully be quite rare, and such tendencies might be kept in check through social pressures or legal protections. Still, it is conceivable that new technological capabilities will multiply the potential harm caused by sadistic acts.

Agential s-risks might also arise when people harbor strong feelings of hatred towards others. One relevant factor is the human tendency to form tribal identities and divide the world into an ingroup and an outgroup. In extreme cases, such tribalism spirals into a desire to harm the other side as much as possible. History features many well-known examples of atrocities committed against those that belong to the "wrong" religion, ethnic group, or political ideology. Similar dynamics could unfold on an astronomical scale in the future.

Another theme is *retributivism*: seeking vengeance for actual or perceived wrongdoing by others. A concrete example is excessive criminal punishment, as evidenced by historical and contemporary penal systems that have inflicted extraordinarily cruel forms of punishment.[30]

These different themes can overlap or take place as part of an escalating conflict. For instance, large-scale warfare or terrorism involving advanced technology might amount to an s-risk. This is both

[30] See Andrews, 2013; Hajjar, 2013.

because of the potential suffering of the combatants themselves and because extreme conflict tends to reinforce negative dynamics such as sadism, tribalism, and retributivism. War often brings out the worst in people. It is also conceivable that agents would, as part of an escalating conflict or war, make threats to deliberately bring about worst-case outcomes in an attempt to force the other side to yield.

A key factor that exacerbates agential s-risks is the presence of malevolent personality traits (like narcissism, psychopathy, or sadism) in powerful individuals.[31] A case in point is the great harm caused by totalitarian dictators like Hitler or Stalin in the 20th century. (I will later return to a more detailed discussion of the concept of malevolence, its scientific basis, and possible steps to prevent malevolent leaders.)

Natural s-risks

The categories of incidental and agential s-risks do not capture all possible scenarios. Suffering could also occur "naturally" without any powerful agents being involved. Consider, for instance, the suffering of animals living in nature. Wild animals are rarely at the forefront of our minds, but they actually constitute the vast majority of sentient beings on Earth.[32] While many people tend to view nature as idyllic, the reality is that animals living in nature are subject to hunger, injuries, conflicts, painful diseases, predation, and other serious harms.[33]

I will use the term **natural s-risks** to refer to the possibility that such "natural" suffering takes place (or will take place in the future) on an astronomical scale.[34] For instance, it would be a natural s-risk if wild animal suffering were common on many planets, or if it eventually spreads throughout the cosmos, rather than remaining limited to Earth. (If human civilisation were to spread wild animal

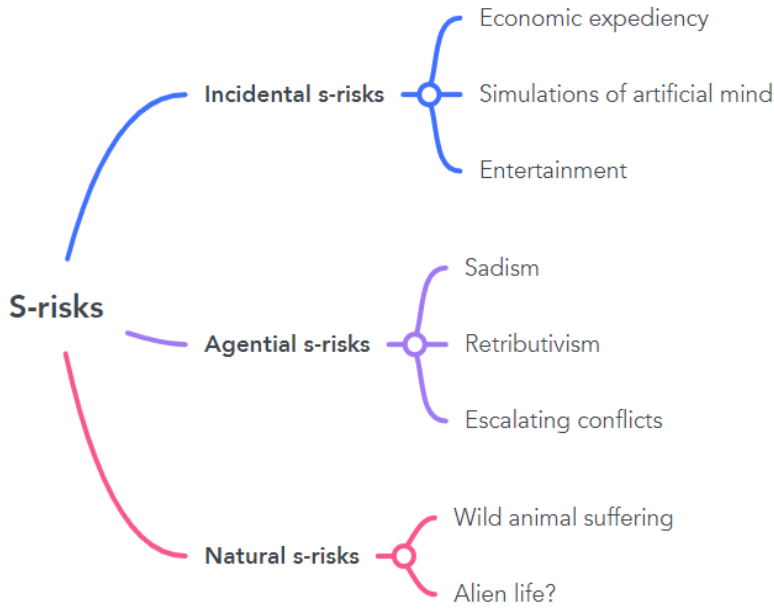
[31] See Althaus & Baumann, 2020.

[32] See Tomasik, 2009; Bar-On et al., 2018.

[33] See Horta, 2017.

[34] If such natural suffering is already taking place now, rather than coming into existence in the future, then it is perhaps inaccurate to call this a "risk". But for simplicity, I will still use the s-risk terminology to refer to the "risk" that we won't do what we could do to reduce such suffering.

suffering throughout the universe, e.g. as part of the process of terraforming other planets, that would count as an incidental s-risk.)



I have so far bracketed the question of how likely or severe the different types of s-risks are. This is partly because the distribution of expected future suffering is an open question that is subject to great uncertainty. Nevertheless, natural s-risks generally seem less worrisome than the other categories. The available evidence — such as the non-observation of any alien lifeforms — does not suggest that wild animal suffering (or anything comparable) is common in the universe. And there is also not much reason, as far as I can tell, to think that a new source of natural suffering (of astronomical proportions) will come into existence in the future.

Other classifications

The classification I have presented so far is based on the mechanisms through which astronomical future suffering might come about. Of course, this is not the only way to classify s-risks.

Another helpful category are **known** and **unknown** s-risks. A known s-risk is a scenario that we can already conceive of at this point. Yet our imagination is often limited. Unknown s-risks may emerge in scenarios that we never thought of or cannot even comprehend (akin to the atomic bomb from a medieval perspective). It is possible that unanticipated mechanisms will lead to large amounts of incidental suffering ("unknown incidental s-risks"), that future agents will have unanticipated reasons to deliberately cause harm ("unknown agential s-risks"), or that new insights reveal that, despite appearances, our universe contains astronomical amounts of natural suffering ("unknown natural s-risks").

We can also distinguish s-risks by the type of sentient beings who are affected. There are s-risks that affect humans, s-risks that affect nonhuman animals, and s-risks that affect artificial minds. The examples I have discussed were primarily focused on the two latter categories. This is because s-risks affecting non-humans are likely to be more neglected — although this does not mean that s-risks that mainly affect humans are unimportant.

Finally, it can be useful to distinguish between **influenceable** and **non-influenceable** s-risks. An s-risk is influenceable if it is possible in principle to do something to prevent it, even if this may be (for whatever reason) difficult in practice. For obvious reasons, I focus entirely on influenceable s-risks, and all examples I discussed so far are influenceable. Non-influenceable astronomical suffering, e.g. in a non-reachable part of our universe, is lamentable but not worth our attention or effort.[35]

[35] As a result of the expansion of the universe, distant objects recede from an observer, and the laws of physics therefore impose a limit on the reachable amount of space. It is worth noting, though, that some decision theories suggest that we may still have an effect on causally disconnected regions if there are correlations between the relevant decision-making processes. See Oesterheld, 2017.

Part II

Should we focus on s-risks?

Introduction

Among the myriad ways to do good, why should reducing s-risks be a main priority? You might think that many of the examples I have given seem far-fetched, or that the discussion of such future risks boils down to unfounded speculation. So why should we take this seriously?

A common measure of the seriousness of a risk is its *expected value* (EV).[36] The EV is the product of the scope of a risk and its probability of occurrence. For example, suppose that disease *A* afflicts one individual, while disease *B* afflicts 20 individuals (equally badly) with a probability of 10% (and none otherwise). According to the principle of EV maximisation, we should (all else equal) prioritise preventing disease *B* because its prevention has a higher EV (20 individuals * 10% probability = 2 afflicted individuals in expectation, rather than 1).

The scope of s-risks is, by definition, astronomical. It is difficult to intuitively grasp the staggering scale of cosmic outcomes. Compared to

[36] For a defense of the idea that we should optimise for expected value (in the context of altruistic efforts), see Tomasik, 2007.

present-day sources of suffering like factory farming or wild animal suffering, s-risks are not just twice as large or even 10 times as large. Instead, they are larger by a factor of thousands or millions, perhaps even billions. Thus, unless the probability of occurrence is vanishingly small, the expected value of s-risks is enormous. To avoid that conclusion, one would need to be extremely confident that s-risks will not happen.

I will discuss reasons for optimism and pessimism about the likelihood of s-risks in detail later. For purposes of this simple argument, it suffices to note that one can hardly justify an extremely low probability. To give a specific number, I claim that the probability of an s-risk materialising is **not less than 1 in 1000**, in light of contemporary analogues (like factory farming) and a range of plausible mechanisms for how s-risks could come about. This lower bound and the vast scope of s-risks suggest (in the EV framework) that averting s-risks should be a priority.

Of course, this simple analysis is not entirely satisfactory and we should dig deeper. I will break down the case for a focus on s-risks into three underlying views:

1. **Long-term focus:** We should mostly focus on improving the long-term future, rather than primarily trying to help those alive now or in the near future.
2. **Suffering focus:** We should prioritise averting severe suffering relative to other goals such as creating a utopian future for humanity. This can be justified on moral or on empirical grounds.
3. **Worst-case focus:** The most effective way to reduce expected suffering in the long term is to focus on preventing particularly bad outcomes that contain a lot of suffering.

I will address each of these points in turn. My goal is to provide an even-handed overview of the ideas that underpin a focus on s-risks, as well as possible reasons to reject such a focus in favor of other priorities.

It is worth noting that the endorsement of these views is gradual rather than binary. We should ask *to what extent* we believe in each of the underlying views, and to what extent we prioritise s-risks as a result. Similarly, a person who only endorses two of the three above points may still be concerned about s-risks to a significant degree.

CHAPTER THREE

Should we focus on the long-term future?

Most people who want to improve the world tend to focus on helping individuals alive today. We more readily empathize with the suffering of those living now than those who will exist a thousand or a million years in the future.

Yet what reasons do we have to discount the interests of future individuals merely because they live at a different time? It seems that the time at which someone exists is not relevant to their moral status. From an impartial perspective, the fact that we live in a certain time does not grant individuals living at this time any special ethical significance. So disregarding the suffering (or other interests) of an individual because of the time they live in would be akin to denying them equal moral status because of other contingent factors, such as the place they live in. This seems arbitrary. It can be said to be a type of discrimination, similar to disregard for those who do not share our skin color or gender, or for those who belong to a different species.

But we should distinguish this notion of *time impartiality* — that suffering matters equally regardless of when it is experienced — from the claim that long-term consequences of our actions should, in practice, guide our decisions. I will refer to this as a *practical* long-term focus.[37] Time impartiality lends support to, but does not necessarily

[37] Some scholars use the term *longtermism* to describe the view that the most important determinant of the value of our actions today is how those actions affect the very long-run future. See Greaves & MacAskill, 2019.

imply, a practical long-term focus. This is because the latter also depends on the number of beings in the future and the extent to which we are able to help them.

I think the philosophical case for time impartiality is overwhelming, so I will primarily discuss questions relating to the practical side.[38]

The future could be vast

The main argument for a practical long-term focus is the potentially vast number of future beings. On cosmic timescales, our lifetime is but an instant compared to the millions and billions of years to come. To give some numbers: according to current cosmological models, the universe is 13.8 billion years old[39] and the remaining lifespan of the sun (before it turns into a red giant) is estimated at 5 billion years.[40] The oldest fossil evidence of modern Homo sapiens dates to around 300,000 years ago (or 0.0003 billion years).[41] So the individuals who are alive today, and who will live in the coming decades, are vastly outnumbered by those who will live in the centuries, millennia, and ages to come (bracketing the possibility of imminent extinction).

In addition to being *long*, the future could be *big* in the sense of containing vast numbers of sentient individuals per generation. We have already encountered the idea that an expansion into space could raise the stakes. Space colonisation could potentially multiply the number of sentient beings by many orders of magnitude.

It is, of course, highly uncertain whether the future will be long, big, both, or neither. But the point still stands. Akin to the previous argument regarding the EV of s-risks, we can hardly be certain that the future will be neither long nor big, and this is enough to establish that

---

[38] For an argument in favour of time impartiality, see Cowen & Parfit, 1992.

[39] Planck Collaboration, 2020.

[40] See Gesicki et al., 2018. Of course, these estimates are highly uncertain, and it is not entirely clear what the relevant endpoint is. But it is clear that cosmic timescales are extremely long — and even just 1 million years is enough for the purposes of this argument.

[41] Handwerk, 2021.

future beings vastly outnumber present-day beings *in expectation.*[42]

Given a moral view that urges us to help others as effectively as possible, this strongly suggests that we should focus on how our actions can benefit future beings.

Influencing the long-term future is difficult

The most common argument against a practical long-term focus is the difficulty of influencing the distant future.[43] The idea that we should optimise for long-term impact, rather than short- or medium-term impact, only makes sense if we can actually do something now to reliably improve the long-term future.[44]

Unfortunately, our impact on the distant future is less predictable than our shorter-term impact. We face great uncertainty over what the future will look like.[45] The world is vast, our knowledge base is limited, and the lack of reliable feedback loops hampers a trial-and-error approach. (This is not to say that further research is futile; in fact, I will later argue that research is one of the most important activities at this point.)

Another challenge is that most of what we can hope to affect now can, and likely will, be changed by later decisions. And future decision-makers may be in a better position to solve future problems than we are. For instance, future people will likely know better which s-risks are most serious, which might give them the upper hand in finding effective interventions. So perhaps reducing suffering in the short term is our comparative advantage.

But this approach of "punting to the future" is risky. It might be

---

[42] One possible justification for a very low probability of a long or big future is a so-called *anthropic penalty* that is linear in the size of the future. If almost all humans exist in the distant future, it is (so the argument goes) unlikely that we would find ourselves in our current position. Hypotheses involving extremely large numbers of future individuals should thus be discounted. See Bostrom, 2013 and Yudkowsky, 2013 for more on this complex topic.

[43] See Tarsney, 2019.

[44] See also Harris, 2019 for some historical case studies that can serve as evidence on how tractable it is to change the course of history.

[45] For more details on this point, see Chapter 9 in Vinding, 2020a.

too late to start thinking about s-risks when they already start to materialise. Without sufficient foresight and caution, society may already be on a trajectory that ultimately leads to a worst-case outcome. And even if future actors are *able* to prevent s-risks, it is not clear whether they will *care enough* to do so. Therefore, we can still do useful things now to ensure both sufficient ability and motivation to reduce s-risks. For example, we can establish a field of research that future people can draw on, and we can foster a community of people who are interested in the topic. (I will elaborate on why these interventions are promising in Chapter 11.)

The hurdles we face when attempting to improve the long-term future are significant, but not insurmountable. It seems hard to argue that the endeavour is entirely futile. Proponents of the practical long-term focus could therefore argue that the challenges of influencing the long-term future do not counterbalance the vast number of future individuals. After all, future individuals outnumber present-day individuals by many orders of magnitude (in expectation). So even if great uncertainty reduces our ability to benefit beings in the distant future by 99 percent (compared to our ability to help existing beings), the expected value framework may still favour focusing on the long-term future, because the scope is larger by more than a factor of 100.

But the following argument can potentially counterbalance the vastness of the future. If the future is long or big, then the outcome is likely to be determined by the decisions of a vast number of individuals. Unless we are in a special situation, our influence might be diluted roughly in proportion to the number of future individuals. The longer and bigger the future, the higher the stakes — but the lower our influence on the long-term outcome. At first glance, these effects cancel each other out.[46]

This argument is abstract and simplistic, but it reveals an interesting observation. The fact that the future could be long or big is not, in itself, sufficient to establish that we should focus on the long term. We should also ask whether we are in a unique position with a lot of leverage over the future.[47]

---

[46] See Baumann, 2019a for more details.

[47] While being in a special position would lend additional support to a

Are we in a good position for long-term influence?

Some scholars have discussed the "hinge of history" hypothesis — that is, the idea that events in this century will have a disproportionate influence on the long-term future.[48] One reason is that a pivotal event could occur in the foreseeable future — such as the aforementioned development of smarter-than-human AI. Such unusual events could lead to a "lock-in" of certain values and power structures, resulting in a steady state that determines everything that happens afterwards.[49] This would avoid the dilution of our long-term impact and thus endow our generation with unusual leverage over the long-term future.

But what evidence can justify the extraordinary claim that we live at a pivotal time? We should be wary of narratives that feed our desire to be special in some way. If history is any guide, it seems more likely that the foreseeable future will entail a gradual, often chaotic evolution of values, institutions, power structures, and many other features — with no discernible endpoint.

On the other hand, we do not need to fully endorse the hinge of history hypothesis to defend a practical long-term focus. And we do have grounds to think that our time has substantial leverage over the long-term future. The simplest reason for this is that we are *early*. The observation that we are still on a single planet that could potentially originate a vast cosmic civilisation suggests that we might indeed be in a unique position. If a long or big future happens, then almost all individuals will live in the future, which means that we can influence them but they cannot influence us.

---

practical long-term focus, it is not strictly necessary — i.e., focusing on the long term could still be best even if our generation is not "special".

[48] See MacAskill, 2020 for an overview.

[49] However, one should not think that powerful AI or other pivotal events would automatically result in a lock-in. It is not clear whether or in what way a lock-in would be reached in these scenarios, or what exactly lock-in means. I am also bracketing the question of whether such a lock-in would be desirable. For some discussion of these vexing issues, see Tomasik, 2017 and Baumann, 2019b.

A less abstract argument is the high pace of technological progress, social change, and economic growth in the last centuries and in the present century. In fact, it is physically impossible for economic growth to continue at the current levels for more than a few thousand years.[50] This lends support to the view that our time — as well as the preceding and following centuries — is particularly influential, although the relationship between economic growth and the degree of influence over the long-term future is complex.

Finding a middle ground

This brief overview of relevant considerations has merely scratched the surface. All things considered, the case for primarily focusing on the long term in our efforts to improve the world seems strong, and I will mostly adopt this view in the remainder of the book.

Yet we must not forget about the terrible suffering that takes place in the here and now. We need to strike a balance between the urgency of reducing ongoing suffering and the goal of improving the long-term future.[51] After all, these are not diametrically opposed aims. So the fact that future individuals vastly outnumber present-day individuals does not mean that we should disregard the latter altogether.

A plausible middle ground is a focus on improving the state of civilisation one or two centuries from now, which will also likely translate to better outcomes in the very long term.

In particular, this means that we should seize low-cost opportunities to support efforts to alleviate ongoing suffering. A case in point is the previously mentioned issue of wild animal suffering. Since few people have worked on improving the welfare of animals living in nature, it is likely that there are still relatively easy and cheap ways to make progress — even if only by raising awareness and doing further research.[52]

Likewise, it is without doubt a worthy cause to end what is one of

---

[50] Vinding, 2017.

[51] A similar conclusion, along with many additional arguments, is found in Tomasik, 2015a.

[52] To learn more about wild animal suffering, I recommend Section 11.2 in Vinding, 2020a and Animal Ethics, 2020.

the largest moral catastrophes of our time: the exploitation of non-human animals and the billionfold suffering it entails.[53]

I will discuss interventions to improve the long-term future in Part 3. I will also come back to the question of why and how animal advocacy can be helpful from a long-term perspective.

[53] For more details on this point, see Vinding, 2015; Horta, 2022.

CHAPTER FOUR

Should we focus on reducing suffering?

To what degree should we prioritise the avoidance of (severe) suffering? The answer depends both on judgment calls in moral philosophy and on empirical beliefs regarding the expected quality and quantity of future lives.

The ethical part is about how much weight we give to the reduction of suffering (or other harms) compared to other goals. The range of possible priorities for the long-term future includes ensuring human survival, creating additional happy lives, increasing the happiness of individuals that are already well-off, and increasing the probability of a utopian future.[54] This raises the thorny question of how much we value and prioritise these different goals, and how we can measure and compare suffering and happiness to begin with.[55]

The empirical part concerns our degree of optimism or pessimism about how much suffering the future contains in expectation. The more optimistic one's view of the future is, the more suffering-focused one's moral view needs to be to prioritise s-risk reduction.[56] Conversely, if a

---

[54] It is worth noting, though, that many things other than suffering would still be important in a suffering-focused view, if only because of their role in enabling us to reduce suffering. For more details on this point, see Ajantaival, 2021a.

[55] See Knutsson, 2016.

[56] Of course, this also depends on other factors, such as the tractability and neglectedness of work on s-risks. I will discuss these other factors later.

large-scale moral catastrophe appears likely to occur in the future, then a large range of moral views will assign great priority to mitigating such an outcome.[57]

Suffering-focused ethics

Proponents of *suffering-focused ethics* argue that the reduction of suffering is of primary moral importance.[58] It stands to reason that this family of moral views, in conjunction with a long-term focus, tends to support a focus on s-risks.

There are many varieties of suffering-focused ethics, which can be supported by many lines of argument. One class of views endorses an asymmetry in the moral importance of creating happy lives compared to preventing miserable lives. In this view, failing to create a happy life does not constitute a bad or a moral wrong on par with the creation of a miserable life.[59] The sentiment is expressed in Jan Narveson's famous dictum: "We are in favour of making people happy, but neutral about making happy people".[60]

Another common view relates to the idea that (severe) suffering carries a unique moral urgency and a corresponding duty to help. By contrast, the mere absence of happiness or other purported goods carries (so the argument goes) no such urgency.[61]

---

[57] Althaus, 2018 suggests that we can quantify this by considering the *normative suffering-to-happiness trade ratio (NSR)*, which measures how we would trade off suffering and happiness in theory, and the *expected suffering-to-happiness ratio* (*ESR*), which measures the (relative) amounts of suffering and happiness we expect in the future. If the product of NSR and ESR is high – either because of a normative emphasis on suffering (high NSR) or pessimistic views about the future (high ESR) – it's plausible to focus on s-risk-reduction. Those who emphasize happiness (low NSR) or are optimistic about the future (low ESR) will tend to focus on other priorities, such as extinction risk reduction.

[58] See Mayerfeld, 1999; Gloor, 2016b; Vinding, 2020a.

[59] For more on this, see Section 1 in Gloor, 2016b and Section 1.1 in Vinding, 2020a.

[60] Narveson, 1973, p. 80.

[61] For more information, see Chapter 6 in Mayerfeld, 1999; Gloor, 2016b,

An example of an idea that can support this view is antifrustrationism, which says that "we don't do any good by creating satisfied extra preferences" and that "you can't do any better than having no frustration".[62] Another idea is that happiness consists in tranquillity or peace of mind.[63] Or one could argue that pleasure is undisturbed affection or the absence of distress, irritation, pain, worry, and so on, and that pleasure cannot increase from that point.[64]

Another common theme is the unique badness of extreme suffering in particular. The worst forms of suffering are seen as much worse than the best states of happiness are good. Extreme suffering can therefore not be "outweighed", or at least not easily, according to this family of views.[65]

Of course, suffering-focused views have also been the subject of criticism.[66] Other value systems, such as classical utilitarianism, contend that creating independent positive goods, like happiness, is of comparable urgency and can morally "outweigh" or "cancel out" any amount of suffering. Depending on how large the required amount of positive goods is, such views might consider the realisation of vast possibilities of a utopian future more important than s-risk reduction.[67]

It is worth noting, though, that a focus on s-risks need not be predicated on views according to which we should exclusively reduce suffering. After all, there is also a broad range of moderately suffering-focused or pluralistic views. And views that hold that (severe) suffering can readily be outweighed would usually still consider the reduction of s-risks valuable, even if it is not a top priority.

Section 3; Gloor, 2017; Vinding, 2020a, Section 1.4. and Chapter 2; Vinding, 2022c.

62 Fehige, 1998, p. 518 and 523.

63 Beiser, 2016, p. 208.

64 See Knutsson, 2022, Section 2, and the references therein.

65 For a more elaborate defense of this, see Chapter 4 in Vinding, 2020a, as well as Section 2 in Gloor, 2016b.

66 See e.g. Chapter 8 in Vinding, 2020a for common objections and possible replies.

67 Cf. Bostrom, 2003.

A detailed review of the pertinent subfields of moral philosophy is beyond the scope of this book. For a more elaborate discussion, I refer the reader to Magnus Vinding's book *Suffering-Focused Ethics: Defense and Implications* (2020) and Jamie Mayerfeld's *Suffering and Moral Responsibility* (1999).[68]

S-risks are not extremely unlikely

A key question on the empirical side is how likely s-risks are.[69] Those who are confident that the future will be bright will tend to give less weight to dystopian scenarios.

A common argument for optimism is that future technology will render it easier to achieve desired outcomes without causing suffering.[70] Since people care at least a tiny bit about avoiding suffering, existing levels of concern will likely be sufficient as soon as the costs of avoiding suffering become very small. An example of this dynamic could be meat that is grown in cell cultures, without the need to raise or slaughter animals. Such technologies might render animal farming obsolete in the future, and may thus prevent suffering without requiring everyone to care deeply about the plight of non-human animals.

While this argument has some merit, it is not airtight. For instance, it does not apply to agential s-risks, which might be exacerbated by powerful technology combined with a relatively low level of concern for suffering. And given our great uncertainty about the future, such abstract arguments are in any case only a weak reason for optimism.

A related argument refers to the empirical observation of positive

---

[68] I also recommend Teo Ajantaival's sequence on minimalist axiologies. See Ajantaival, 2021a, 2021b and 2022.

[69] More precisely, the relevant question is how effectively we can use marginal resources to reduce future suffering, compared to how readily we can achieve positive goods or other goals. This question also encompasses the relative tractability of interventions to reduce s-risk and interventions to achieve other goals. But for reasons of simplicity, I will focus only on the likelihood of s-risks in this section.

[70] See West, 2017.

trends over the last centuries, e.g. in terms of declining violence,[71] improved health,[72] and many other metrics. It is undoubtedly true that the world has improved in some ways, but this perspective still seems anthropocentric. When taking into account the rapid increase in the number of farmed animals since the industrial revolution,[73] it is far less clear whether the world is getting better or worse over time. Human progress may have (incidentally) multiplied animal suffering, and the number of animals slaughtered is still rising.[74]

This is not to say that I endorse a highly pessimistic view of the future. Instead, the point is merely that a high level of confidence in an optimistic view is not warranted. And unfortunately, there are several reasons to believe that the probability of s-risks is not vanishingly small.

First, s-risks can materialise in many ways, as discussed in Chapter 2. It is hard to predict the future and we can only imagine a limited range of scenarios. So even if any particular scenario that we can conceive of seems highly unlikely, the probability of *some* (perhaps unknown) s-risk may still be non-negligible.

Second, s-risks may seem speculative at first, but the underlying assumptions are quite plausible. Barring globally destabilizing events, we have no reason to expect that technological progress will soon come to a halt. It is not speculative to argue that more powerful technology will raise the stakes, for better or worse. And it is also plausible that those in power will continue to show insufficient concern for the suffering of less powerful beings — akin to the current lack of concern for the suffering of nonhuman animals or disadvantaged humans.

Third, historical precedents exist. It is tempting to see humanity on

---

[71] Pinker, 2011.

[72] Roser et al., 2013.

[73] This is especially true when taking into account aquatic animal agriculture, which has expanded massively in scale. See Ritchie, 2019.

[74] See Ritchie and Roser, 2017, Section "Number of animals slaughtered". Note that this primarily relates to anthropogenic harm from animal exploitation. An overall assessment of whether the situation of animals has improved or worsened is complex, especially when taking into account animals living in nature. After all, the overall impact of human civilisation on wild animals has not been studied in depth.

a path towards ever greater moral progress, but severe moral regress can also occur. A case in point is the rise of fascism in the 20th century, which was unimaginable to many contemporary observers.[75] We also have a mixed track record regarding the responsible use of new technologies — from novel tools for brutal torture to chemical weapons and nuclear bombs. Likewise, the situation of nonhuman beings has worsened in many ways with the advent of factory farming. So we can hardly be certain that future technological developments will be handled with appropriate care and consideration.

To clarify, working on s-risks does not require a pessimistic view of the future. The above arguments are consistent with believing that technology can also benefit us, that we could use it to reduce rather than increase suffering, or that our quality of life may improve to unprecedented levels. To be concerned about s-risks, it is sufficient to believe that the probability of a bad outcome is *not negligible*, which is consistent with believing that a utopian future is *also* quite possible.[76]

A focus on suffering is at least a plausible perspective

Plausible ethical arguments and empirical considerations lend support to a focus on suffering.[77] And while not all combinations of moral and empirical views consider the reduction of suffering the foremost priority in practice, it is uncontroversial that preventing or alleviating severe suffering is of great importance. Virtually everyone would agree that such suffering should, all else equal, be avoided. A

---

[75] See Vinding, 2022f for more details and additional historical examples of moral regress, as well as possible implication in terms of optimism or pessimism about the future.

[76] We should also distinguish between the average level of well-being of future individuals and the *absolute* amount of suffering, especially if the future will contain vastly larger populations. The absolute amount of suffering can increase due to a larger population size, even if the average level of well-being improves or remains the same.

[77] Of course, this brief overview is not necessarily sufficient to establish that the reduction of suffering is the *top* priority in terms of how we should use marginal resources. This is a much harder argument to make, and I refer the reader to Vinding, 2020a for a more elaborate case for prioritising suffering.

focus on suffering is particularly widely accepted if conceived of as a (strong) *component* in a moral view, rather than being exhaustive.

This is not to deny that there is profound disagreement. People have different priorities, and this will remain true in the foreseeable future. This is a reason to seek to foster cooperation between people who endorse different value systems.[78] Instead of narrowly pursuing our own ethical aims in potential conflict with the aims of others, we have strong reason to seek common ground and to embrace projects that are valuable from many perspectives.[79] Since avoiding vast amounts of suffering is a common interest of many value systems, s-risk reduction is a good candidate for a shared aim that people with different values can agree on.

[78] For more details on this, see Chapter 10 in Vinding, 2020a, as well as Tomasik, 2013a, Ord, 2015 and Vinding, 2020b.

[79] See Baumann, 2020a for an overview of what this common ground could look like.

CHAPTER FIVE

Should we focus on worst-case outcomes?

The combination of a long-term focus and a suffering focus implies that we should work towards reducing future suffering. A focus on s-risks entails the additional belief that guarding against particularly bad outcomes is (in expectation) the most effective way to reduce future suffering. I will refer to this as a **worst-case focus**. Without this belief, the concept of s-risks might not add much value, as it would be simpler to just talk about reducing future suffering.

Consider a toy model with three possible futures:[80]

A. A future with no suffering, e.g. due to improved moral standards or advanced technology that makes it possible to abolish suffering altogether.

B. A future that contains similar amounts of suffering as now, in which some people live in abject poverty, factory farms (or future equivalents thereof) continue to exist, and wild animal suffering is never tackled to a sufficient degree.

C. A future that contains significant levels of suffering on an astronomical scale — i.e., a future in which an s-risk materialises.

The worst-case focus rests on the claim that the difference between B and C is much greater than the difference between A and B, so that preventing C is most important — even taking into account a lower probability of C. By contrast, a rejection of the worst-case focus would imply a belief that most expected suffering lies in futures that are

[80] Adapted from Gloor, 2018.

moderately bad, but not so bad as to qualify as s-risks. In this case, it might make more sense to increase the probability that suffering is abolished altogether.[81]

Is future suffering heavy-tailed?

Why should we believe that most expected suffering in the future stems from worst-case outcomes?

Many phenomena, from the size of earthquakes to the number of fatalities in wars, have been modeled using a power-law distribution.[82] The power law is an example of a *heavy-tailed* distribution in which extreme outliers are common.[83] In other words, most casualties of war are concentrated in a relatively small number of the bloodiest wars, and most of the overall damage from earthquakes is due to the most extreme ones. This tentatively suggests that a similar pattern might hold for future suffering — i.e., that a large fraction of (expected) suffering might be concentrated in the most extreme s-risks.[84]

This is only weak evidence because it is not clear how comparable all these different phenomena are. And many other distributions are less heavy-tailed. Consider, for instance, the distribution of income in the United States. Income is highly unequal, and the top 1% of earners make far more than others. Yet they still make up less than 20% of the total income.[85] This is not to say that the existing degree of economic inequality is unproblematic, especially since wealth is more skewed than income. My point is merely that this is an example where the outliers do not completely dominate the distribution, or at least not as strongly as they do in the case of wars or earthquakes.

A reason to think that s-risks are a heavy-tailed phenomenon is that the scale of future suffering could, as mentioned before, vary by

---

[81] This approach has been termed the "abolitionist project" or the "hedonistic imperative". See Pearce, 1995 for a defense and Vinding, 2021 for a critique of the abolitionist project (as a top priority).

[82] Becerra et al., 2012.

[83] For more details on the underlying math, see Newman, 2005.

[84] See Baumann, 2021 for more details.

[85] See Figure 5 in Donovan et al., 2016.

many orders of magnitude (e.g. due to the vastness of a potential intergalactic civilisation). This difference of many orders of magnitude is difficult to grasp intuitively, as our minds did not evolve to deal with such large numbers. However, quantities like income or war casualties can also vary by many orders of magnitude, so this is perhaps not unique to s-risks.

All things considered, a worst-case focus seems at least plausible, though it remains unclear to what degree we should expect future suffering to be concentrated in the most extreme outcomes. It is also important to keep in mind the gradual nature of the worst-case focus. While many real-world phenomena are somewhat heavy-tailed, they are almost never *extremely* skewed, such that, say, the most extreme 0.1% account for 99% of the distribution. So we should perhaps focus on, say, the worst 10% of outcomes, rather than the worst 0.1%.

Reasons to avoid a narrow focus on just a few scenarios

It is important to distinguish the worst-case focus from the notion that we should concentrate on a small number of specific worst-case scenarios.[86] I will refer to the latter as a *narrow* focus. At first glance, the worst-case focus appears to imply such a narrow focus. But a heavy-tailed distribution does not necessarily mean that most (expected) future suffering would come from just a few specific sources.[87]

After all, the most extreme 1% of a distribution can still be very diverse. As a case in point, the richest 1% still vary a lot in terms of how they acquired their wealth or in terms of their political attitudes. And they are also a large group in absolute terms (e.g., one percent of the US adult population is still more than 2.5 million people). Similarly, the worst 1% of possible futures could include a substantial number of diverse scenarios and sources of suffering.

---

[86] Most points in this section are inspired by Vinding, 2020c.

[87] I am bracketing the mathematical complication that the expected value of such distributions often does not even exist, which poses a challenge to the expected value framework.

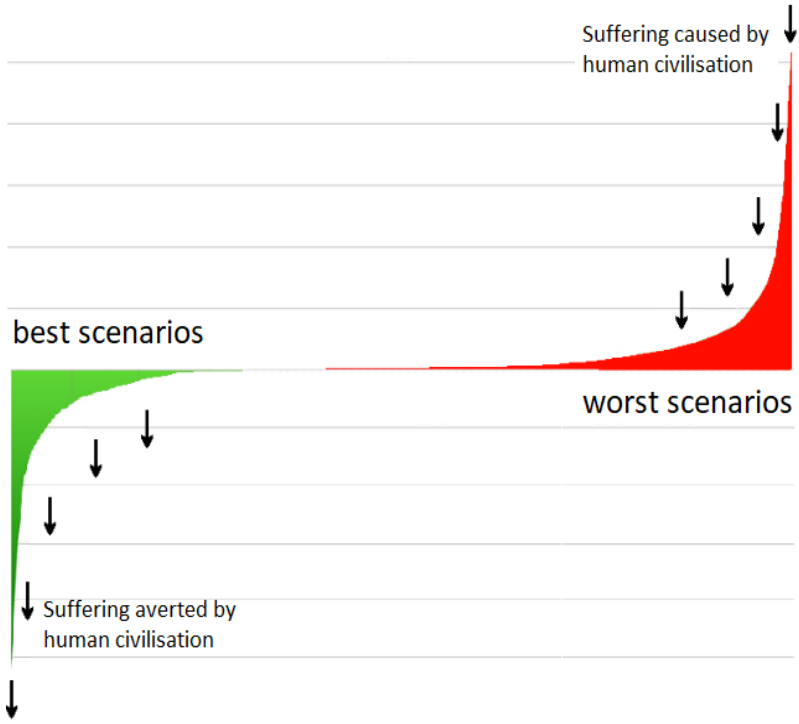
A narrow focus is also risky because of great uncertainty. Even if most suffering occurs in a limited number of specific worst-case scenarios, we might still be mistaken in our best guesses as to what those scenarios are. Indeed, if our guesses rely on little more than speculation, it seems *very likely* that the worst scenarios are among those to which we currently give little weight. In light of great uncertainty, a narrow focus on a few specific scenarios seems suboptimal.

Even if a small number of scenarios make up a large fraction of expected suffering, it could still be better on balance to address a broader range of risks. This is the case if more targeted interventions against the worst risks are not available or not significantly more effective than broader interventions that are helpful with respect to a wider range of risks.

It is worth noting, though, that the discussion above is about what people who are interested in reducing s-risks should do as a collective. It can still make sense for individual people to specialise on a narrow set of scenarios, as long as our efforts to reduce s-risks collectively cover a sufficiently broad set of risks.

Finally, it would be misguided to only consider the "negative tail" of worst-case outcomes that are caused by human civilisation. In the case of natural s-risks or astronomical suffering caused by alien civilisations, a "positive tail" could arise from potential opportunities to prevent those s-risks.[88]

[88] While the universe appears empty (cf. Sandberg, 2018), a recent model of "grabby aliens" suggests that space will likely be colonised by alien civilisations, regardless of whether humanity will eventually do so — cf. Hanson et al., 2021; Cook, 2022. Additionally, it is conceivable (albeit speculative) that our actions could affect other parts of the multiverse through correlated decision-making. See Oesterheld, 2017 for more details.



The figure above illustrates how the impact distribution could be *double-tailed*, i.e. heavy-tailed in both directions.[89] It shows percentiles along the x-axis and the amount of suffering that is created or reduced by human civilisation along the y-axis. The red part is the negative tail (suffering caused by human civilisation) and the green part is the positive tail (suffering averted by human civilisation). As symbolised by the arrows, it would be beneficial to reduce the negative tail and to increase the positive tail.

For various reasons, such as the apparent emptiness of the universe, it seems plausible that the negative tail is larger. Still, in light of great uncertainty, we should give some weight to both tails (as illustrated in the figure). This is yet another reason to prefer interventions that are robust across many scenarios.[90]

[89] The figure is taken from Vinding & Baumann, 2021.

[90] See Vinding & Baumann, 2021 for more details on this point.

CHAPTER SIX

Cognitive biases

Modern research has documented that the human mind is prone to a long list of *cognitive biases*: systematic patterns of flawed thinking or mental shortcuts that often cloud our judgment.[91] In the following, I will argue that common biases could cause us to mistakenly dismiss or downplay s-risks. To form an accurate appraisal of s-risks, we need to keep these biases in mind and try to control for their distorting effect.[92]

Wishful thinking

One of the most well-documented biases is wishful thinking: a tendency to believe what we wish were true, rather than what is supported by the available evidence.[93] Wishful thinking is likely to be a significant source of bias against taking s-risks seriously.[94] We dream of a bright future, and the possibility of a vast moral catastrophe

---

[91] See e.g. Tversky and Kahneman, 1974.

[92] It is, of course, doubtful whether mere awareness of a bias suffices to overcome them. However, I will bracket the question of how to best overcome biases as it is beyond the scope of this book.

[93] Bastardi et al, 2011. See also Tomasik, 2015b for an analysis of wishful thinking in the context of present and future suffering.

[94] See also Section 2.1 in Althaus & Gloor, 2016.

disturbs this dream.

This bias is plausibly exacerbated by the unpleasantness of contemplating worst-case outcomes. Preventing s-risks is not a particularly inspiring vision, unlike the prospect of a utopian future or opportunities to help others in the here and now. It can even be emotionally taxing or depressing to devote one's attention to horrific suffering and worst-case futures. So we often prefer to turn a blind eye to others' suffering and to risks of bad outcomes — especially when we ourselves are comfortable and safe.[95]

Psychological research has long documented this *denial* of atrocities, suffering, and other uncomfortable realities.[96] Denial can take many forms, such as not wanting to know about suffering, denying that the suffering is real, downplaying its badness,[97] or disputing any personal responsibility on our part.

Yet another (related but distinct) bias is the just-world fallacy.[98] We like to assume that the world is morally fair and just. If something bad happens to someone, they must have done something to deserve it. This thought is more comforting than is facing a reality that often contains entirely "pointless", horrific suffering.

These psychological tendencies plausibly conspire to produce a bias against taking s-risks seriously.[99] To overcome this bias, we must actively resist the urge to avert our attention from the distressing possibility of astronomical suffering. There is too much at stake.

---

[95] See also Vinding, 2020a, Section 7.2.

[96] See Cohen, 2001 for an overview.

[97] It is important to distinguish severe suffering from far milder emotions that only masquerade as suffering to gain increased attention — such as anger, bitterness, frustration, or mere discomfort. Confusing relatively mild suffering for its more intense manifestations often results in an underestimate of the badness of the worst forms of suffering. For more on these points, see Section 7.4 in Vinding, 2020a and Tomasik, 2006.

[98] Furnham, 2003.

[99] While the above-mentioned phenomena are usually studied in the context of ongoing atrocities or suffering, it seems likely that similar factors apply to discussions of s-risks. We may even be particularly prone to wishful thinking or denial about future risks, because their abstract and uncertain nature can make it emotionally relatively easy to ignore their distressingly large scope. See also Section 7.13 in Vinding, 2020a.

Scope neglect

Scope neglect refers to our inability to fully appreciate the scope of a problem. At the level of our moral cognition, we simply do not "feel" the numbers. Abstract considerations about future suffering do not pull on our emotional heartstrings in the same way that a single identifiable victim does. We may consciously acknowledge that a large-scale moral catastrophe is much worse, but it seems impossible to truly grasp the badness in proportion to the scope.[100]

Scope neglect likely affects our reasoning about s-risks. After all, an exceptionally vast scope is precisely the hallmark of s-risks. Thus, to fully appreciate the vast scale of worst-case futures, we likely need to make a deliberate effort to counter scope neglect.

A related effect is *proportion dominance*: we feel more compelled to help 10 out of 10 individuals than to help 10 out of 1000, even though the absolute impact is the same.[101] This effect likely distorts our intuitions against a focus on alleviating suffering in worst-case outcomes, as we may (wrongly) find it more worthwhile to go from 10 instances of intense suffering to 0 than to go from 1000 to 990.

In general, few people have fully internalised the expected value framework in the context of altruism, even if they agree with it in theory. It is difficult for the human mind to conceive of or comprehend the magnitude and probabilities of s-risks. Likewise, few people have fully internalised the idea that helping future sentient beings is just as important as helping those alive now, or the idea that species membership per se is an ethically irrelevant characteristic. It is thus unsurprising that the consistent application of these principles can result in conclusions that seem counterintuitive, such as prioritising the prevention of worst-case scenarios. Yet if we are confident that the underlying moral and decision-theoretical beliefs are sound, then it is reasonable to follow them to their conclusions.

---

[100] Indeed, people actually seem to care *less* as the numbers increase. See Slovic, 2007.

[101] The effect can occur even at the expense of absolute impact. In a study, people preferred helping 225 out of 300 lives over helping 230 out of 900 lives. See Bartels, 2006.

Belief digitisation

Belief digitisation is the tendency to implicitly assign hypotheses the probabilities of either 1 or 0, rather than accurately taking uncertainty into account.[102] When making predictions or decisions, people tend to implicitly "round down" the probability of an unlikely hypothesis to 0, even when they consciously agree that the hypothesis has non-negligible probability. Conversely, an apparently plausible hypothesis is often taken for granted — i.e., implicitly assigned a probability of 1 — even when there is considerable uncertainty.

This could mean that we may be prone to mistakenly dismiss s-risks because we implicitly round their probability down to 0. This is at odds with valid reasoning about probabilities and expected values, because a small but non-negligible probability can, as outlined earlier, still be enough to justify a focus on s-risks. (For reasons to believe that the probability is not very close to 0, see Chapter 4.)

Apart from the question of whether to prioritise s-risks, belief digitisation can also lead us to focus on an overly narrow range of s-risks. If we are implicitly too confident in certain assumptions (because we round their probability up to 1), we may be inclined to focus on too small a class of scenarios. This is most likely to occur if specific risks or scenarios are particularly salient to us.[103]

More generally, we are often overconfident in our beliefs and thus fail to fully appreciate our uncertainty about the future.[104] One of the most ubiquitous biases that lead to overconfident and inaccurate judgments is confirmation bias: the tendency to seek out and favour information that confirms our beliefs, while ignoring contrary information.[105] In the context of our efforts to reduce suffering, this bias may result in a strong attachment to a particular strategy or cause area, which can in turn compromise our ability to objectively evaluate

---

[102] Johnson et al, 2020.

[103] This is because of the availability heuristic: We tend to estimate the likelihood of events based on how readily available they are in memory. Cf. Kahneman, 2011.

[104] Plous, 1993, p. 219-220.

[105] Mercier & Sperber, 2017.

other causes and strategies.

This highlights the need to remain open-minded and epistemically modest. After all, it is easy to overlook crucial considerations. Examples include animal advocates who give limited consideration to the issue of how animals suffer in nature, or activists who may reduce far more suffering by focusing more on influencing the long-term future. There is no reason to think that we have already uncovered all relevant considerations about the best priorities for reducing suffering.

Concluding thoughts

In light of the biases reviewed above, it comes as no surprise that few people actively work on reducing s-risks.[106] Even people who care about long-term outcomes tend to focus on achieving utopian outcomes or other priorities.

This neglectedness is another strong reason to invest marginal resources in s-risk reduction.[107] Since the topic has not yet been explored in much detail, we might be able to make considerable progress when working on s-risks. Particularly neglected are s-risks that affect nonhuman beings, due to the combination of biases against taking s-risks seriously and the generally insufficient level of moral concern for these sentient beings.

The existence of biases against taking s-risks seriously can therefore be an argument in favour of prioritising s-risks. Of course, such an analysis of potential biases does not in itself give us strong reason to prioritise s-risks. But in light of our great uncertainty, it seems valuable to consider a multitude of different perspectives and arguments.

We should also remain open to the possibility that there may be biases in the opposite direction that cause us to systematically overestimate the scale and likelihood of s-risks. For instance, perhaps extreme scenarios are more interesting to discuss compared to more

---

[106] This is not to say that biases are necessarily the only reason why few people prioritise s-risk reduction.

[107] It is worth noting, though, that some other cause areas, such as reducing wild animal suffering, are also very neglected. Vinding, 2020d outlines various biases against prioritising wild animal suffering.

mundane futures — which may be why dystopian science fiction is popular. Many people may also exhibit a negativity bias, which might contribute to a common tendency to believe in gloomy predictions about impending decline or catastrophe. This is at odds with many metrics that suggest that the world has actually improved over time.[108]

Still, I think it is more plausible, on balance, that we are biased to dismiss s-risks. This is because many of the biases reviewed here seem strong, and most of them are biases against a focus on s-risks, whereas potential biases in favour of a focus on s-risks seem weaker and fewer in number.

This concludes my discussion of reasons for and against a focus on s-risks. All things considered, I think the case for each of the three underlying premises — long-term focus, suffering focus, and worst-case focus — is plausible, but not unassailable. Or, to put it differently, I endorse moderate versions of each of those premises but not necessarily strong ones.

In any case, we should remain open-minded, acknowledge that there are valid arguments for and against a focus on s-risks, and be willing to change our mind if new evidence about the likelihood of s-risks or the (in)feasibility of long-term influence emerges.

[108] See Pinker, 2011.

Part III

How can we best reduce s-risks?

Introduction

I hope that you are, at this point, convinced that s-risk reduction is at least a plausible priority worthy of significant attention (even if you are not convinced that it should be our sole priority). The rest of the book will proceed on that basis and focus on the practical question: **What can we do to avert s-risks?**

We must avoid many pitfalls in the endeavour of reducing s-risks.[109] The first and perhaps most common pitfall is to

[109] Most of the following points are inspired by Chapter 9 in Vinding, 2020a.

underestimate the enormous complexity of the question of how to best reduce s-risks. Given our great uncertainty about the future, it is unrealistic to expect a single, conclusive answer to this question.[110]

And as discussed in the previous chapter, cognitive biases can pull us towards overconfident views or a premature focus on a specific cause area. A particular risk to be mindful of is our tendency to be more concerned with *showing* that we care about an issue, rather than *actually* solving it.[111] One antidote is to cultivate epistemic humility and to actively seek out new information that can inform our approach to reducing s-risks.

Another pitfall is a lack of prudence. We should avoid approaches to s-risk reduction that could easily backfire, e.g. by antagonising those with other priorities. For many reasons, it seems more productive to instead pursue a cooperative approach and to work towards aims that find broad support.[112]

Lastly, we may fail to appreciate that any individual can, in a world of almost 8 billion people, only ever hope to have a limited, *marginal* impact. It is thus misguided to reason as if we can single-handedly steer the future.[113] This relates to the *illusion of control*, which is the tendency for people to overestimate their ability to control events.[114] Given that our individual impact on the trajectory of humanity is limited, it is often not very useful to envision an ideal future and then try to make that happen. A better alternative is to focus on how much impact additional resources could have on the margin, when invested in a particular area.

Yet this does not imply that our impact is small in absolute terms, or that the endeavour of reducing s-risks is futile. I am confident that this is not the case. We do have the power to achieve a lot, especially if we think carefully about how to best reduce s-risks.[115]

---

[110] This has been termed the "silver bullet delusion". See Section 9.4 in Vinding, 2020a.

[111] See Section 8.10 in Vinding, 2022a, as well as Chapter 12 in Simler & Hanson, 2017.

[112] For more details on this, see Chapter 10 in Vinding, 2020a.

[113] This point is made by Hanson, 2014.

[114] Thompson, 1999.

[115] See Section 9.8.1 in Vinding, 2020a for more details on this point.

CHAPTER SEVEN

Risk factors for s-risks

Suppose you want to evaluate how a given intervention would affect s-risks. This is made difficult by the multitude of possible s-risks and by our great uncertainty about the future. Similar to past and contemporary forms of suffering, future suffering will likely result from a variety of issues rather than any single cause. We therefore need to consider measures for s-risk reduction that are comparatively easy to assess, which will then simplify our discussion of specific interventions.

In this chapter, I will introduce several *risk factors* for s-risks. These risk factors are not s-risks in and of themselves, but they significantly increase either the probability or the severity of a very bad outcome. The concept is also used frequently in medicine. For instance, an unbalanced diet or a lack of exercise are not adverse health outcomes in and of themselves, but they are risk factors for a plethora of medical problems, from heart disease to depression.[116]

This framework allows us to give sound advice for a healthy lifestyle without the need to analyse specific diseases. And the resulting conclusions are robust even though the health trajectory of any given individual is highly uncertain.

By analogy, we might not need to know all effects that a given action will have on specific s-risks. If we can identify reliable risk factors, we will be able to derive robust and effective interventions for

[116] See e.g. North et al., 2008 and Lavie et al, 2015.

reducing a broad range of s-risks.

Advanced technology and space colonisation

The simplest risk factor for s-risks is the capacity of human civilisation to create large amounts of suffering in the first place. As discussed in Chapters 1 and 2, many s-risks are only possible in the context of powerful new technologies that give rise to both unprecedented opportunities and unprecedented risks. In particular, the emergence of advanced AI could, due to its unprecedented power, constitute a serious s-risk.[117]

As with the concept of medical risk factors, this does not mean that the emergence of such advanced technologies would necessarily cause an s-risk to materialise.[118] The point is merely that advanced technologies would equip humans with immense power and thereby exacerbate the potential scope of worst-case outcomes.

We should also distinguish between more and less worrisome forms of technological progress. The relevant aspect is the effect that new technologies could have on the overall scale of human civilisation and on the number of (potentially miserable) sentient beings. In particular, some new technologies might make it easier to create large amounts of suffering.

A concrete example of such a technology is the ability to create sentient artificial entities. As I have discussed in Section 1.4, this might result in the exploitation of large numbers of sentient beings, due to our likely insufficient level of moral consideration for artificial minds.

Another key factor is large-scale space colonisation.[119] Due to

---

[117] See Section 1.4. It is also worth noting that advanced AI could not only cause s-risks, but could also help prevent them, as argued in Sotala & Gloor, 2017.

[118] Also, some s-risks would still be possible even if humanity were to halt all technological progress. This includes natural s-risks or potential s-risks caused by alien civilisations. And without advanced technology, we may be unable to do anything about these s-risks.

[119] This term is meant to refer to large-scale settlement across a vast number of planets or even galaxies, not activities like the mere exploration of space or isolated outposts for purposes such as asteroid mining.

advanced AI or other technological breakthroughs, it might become technically and economically viable to expand throughout the universe. This expansion could potentially multiply the total population size of both human and nonhuman beings, resulting in a truly astronomical scope of our civilisation. And without sufficient moral and political progress, this could multiply the amount of suffering entailed by our civilisation.[120]

Thus, space colonisation, even with the best of intentions, poses significant risks. The potential scale of future civilisation is mind-boggling. Astronomers estimate that there are 100-400 billion stars in our galaxy (the Milky Way) alone, and 100-200 billion galaxies throughout the universe.[121] A future moral catastrophe on a galactic or intergalactic scale could therefore exceed Earth-based suffering by many orders of magnitude. By contrast, the amount of suffering is limited if we never expand into space.[122]

Some authors have further argued that space colonisation will by default result in catastrophic outcomes.[123] But this is highly uncertain, and a pessimistic view of space colonisation is not necessary to establish that space colonisation is a risk factor for s-risks.

Is the large-scale colonisation of space a realistic prospect? Evidence on the feasibility of space colonisation is scarce, but preliminary reviews suggest that the obstacles, from microgravity to travel across cosmic distances, are massive yet probably not insurmountable.[124] It also remains unclear what the motivation to colonise other planets would be — considering that other planets are usually extremely inhospitable places when compared to Earth, and we are currently far from running out of available land.[125]

---

[120] If the universe contains or will contain alien civilisations (cf. Hanson et al., 2021), then human expansion into space could also potentially reduce s-risks, as discussed in Section 5.3.

[121] Howell, 2021; Howell & Harvey, 2022.

[122] It is worth noting, though, that the number of beings on Earth could in theory also be very large (albeit still not as large as in a spacefaring civilisation). See Hanson, 2011 for more details.

[123] Deudney, 2020; Torres, 2018a; 2018b.

[124] Beckstead, 2014.

[125] For more details on this, see Baumann, 2020b.

On the other hand, we face great uncertainty about what may or may not happen in the future, especially on long timescales. So we also cannot rule out the possibility that humanity will colonise space. And the scope of a galactic or intergalactic moral catastrophe could be so vast that an expected value framework suggests that we should take the possibility seriously, even if we do not consider large-scale space colonisation to be the most likely scenario.

Lack of adequate s-risk prevention

Human civilisation can likely mitigate most forms of suffering, given sufficient motivation and political will to do so.[126] Therefore, s-risks are far more likely to occur if nobody works to prevent them.

Even a limited degree of moral concern could go a long way towards mitigating s-risks. For instance, if only a small number of people care about preventing an s-risk, they can still try to find a compromise with others to implement low-cost measures that can prevent worst-case outcomes. Such low-cost compromises are likely to be possible for many s-risks. It therefore seems plausible that we should be most worried about futures with little or no efforts to prevent s-risks, and that we should address that apparent bottleneck.

What could cause such a lack of efforts to prevent s-risks? The simplest reason is sheer indifference, especially for s-risks that affect those without any political representation or power. Future decision-makers might be aware of s-risks and be able to avert them, but they might not care enough about the suffering that their decisions cause (or fail to prevent). In particular, a narrow moral circle could result in a disregard of s-risks that affect nonhuman animals or artificial sentience.

Even if there is concern for s-risks, it is possible that the resulting efforts are misguided or ineffective. This could happen for many reasons. For instance, the idea of preventing s-risks or reducing suffering might become associated with controversial political ideas and factions, which could in turn cause a backlash that thwarts progress towards preventing s-risks.

---

[126] However, some s-risks might be hard to prevent even if we collectively want to.

It is also possible that the relevant actors will want to avert an s-risk, but doing so may be impossible due to ineffective political institutions or cooperation problems. Or the relevant actors might lack the foresight to anticipate and address potential s-risks at an early stage — and at a later point, it might be impossible to change course. (Of course, this depends heavily on the specific s-risk in question.)

Conflict and hostility

S-risks are more likely if there is a high degree of **hostility** between future actors, with little or no common ground. It is, of course, not problematic per se if people endorse different perspectives or opinions. However, such divergences can constitute a risk factor for s-risks when combined with a lack of understanding of other perspectives, or intolerance and hostility towards others.

Conflicts can be problematic for several reasons. First, powerful factions or individuals might ride roughshod over the moral concerns of others. This is likely to impede efforts to prevent s-risks (the lack of which is a risk factor, as per the previous section). A future that entails large-scale adversarial dynamics or ruthless competition would likely leave little room for prudent reflection on s-risks or for mutually beneficial compromises. Negative outcomes would be significantly more likely in this case, compared to a future where successful coordination makes it possible to implement countermeasures against potential risks.[127]

Second, hostile relations always carry a risk of escalating conflicts and even outright war between competing factions. It stands to reason that this increases the risk of worst-case outcomes. In particular, some actors might want to intentionally harm others out of hatred, sadism, or vengeance for (real or alleged) harm caused by others (as discussed in Chapter 2). Conflicts and wars tend to exacerbate our worst impulses.

A related risk factor for s-risks is insufficient security against bad actors.[128] Human civilisation contains many different actors, including

[127] See Tomasik, 2013a.

[128] Bostrom, 2019 introduces a similar concept of a *semi-anarchic default condition*, and argues that human society is currently in such a state, and that

some malevolent ones. Such bad actors are usually reined in by norms and laws that prohibit harmful acts, yet this might become difficult in some future scenarios. For instance, in the context of powerful autonomous AI agents or space colonisation, it might become harder or even impossible to stop rogue actors from causing harm on a massive scale.[129]

This is related to the future evolution of the *offense-defense balance*.[130] Military applications of future technological advances could change the offense-defense balance in a way that makes s-risks more likely. A common concern is that strong offensive capabilities would enable a safe first strike, undermining global stability. Yet when it comes to s-risks, it is perhaps even more dangerous to tip the balance in favor of strong defense, since bad actors can no longer be deterred from harmful acts if they enjoy strong defensive advantages.[131]

Malevolent actors

Cruel dictators like Hitler and Stalin were responsible for many of the worst atrocities in human history. But how can we operationalise this notion of "cruel" or "malevolent" actors? A frequently used concept is the "Dark Tetrad", which consists of the following four personality traits:[132]

---

we should attempt to exit this condition by establishing effective global governance and preventive policing. (See also Hanson, 2018 for a reply.)

[129] In the case of powerful autonomous AI, existing laws and institutions may not be directly applicable, and it is not obvious what the replacement could be. In the case of space colonisation, large cosmic distances might constitute an obstacle to effective enforcement of laws or norms — although this is, of course, highly speculative. The point is merely that a breakdown of the rule of law would make s-risks much more likely.

[130] See Garfinkel & Dafoe, 2019 for more details on this concept.

[131] How could strong defensive capabilities come about? One plausible scenario is intergalactic space colonisation with multiple loci of power. It might then be difficult to enforce large-scale prohibitions against harmful acts due to astronomical distances between galaxies or superclusters.

[132] Paulhus, 2014. Note also that the "dark traits" are positively correlated with each other, which is why it makes sense to combine them into a single "Dark Factor" — see Moshagen et al, 2018. This has also been contrasted with

- **Psychopathy** is characterized by persistent antisocial behavior, impaired empathy, callousness, and impulsiveness.
- **Narcissism** involves an inflated sense of one's importance and abilities, an excessive need for admiration, and an obsession with achieving fame or power.
- **Machiavellianism** is characterized by manipulating and deceiving others to further one's own interests, indifference to common norms, and ruthless pursuit of power or wealth.
- **Sadism** is the tendency to derive pleasure from inflicting suffering and pain on others.

Individuals with malevolent traits can pose serious risks if they rise to positions of power.[133] And they often do — after all, the hallmarks of malevolence include strategic ruthlessness and a lust for power. These traits are often an advantage in the struggle for power, especially in fiercely competitive systems.[134]

Malevolent individuals in power can cause a variety of negative outcomes. The aspects that are most relevant to s-risks include an erosion of interpersonal trust and coordination, an increased risk of escalating conflicts and war, and an increased likelihood of reckless behaviour in high-stakes situations.[135]

A concrete pathway to an s-risk is the formation of a global totalitarian regime under a malevolent leader, which could potentially result in a permanent lock-in of ruthless values and power structures. Historical examples of totalitarian regimes (e.g., Nazi Germany or Stalinist Russia) were temporary and localised, but a stable global dictatorship may become possible in the future.[136]

Risks from malevolent actors are exacerbated if those actors have access to advanced technology, such as powerful AI. In the worst case,

---

a "Light Triad" of beneficial traits; see Lukić & Živanović, 2021.

[133] This section largely follows Althaus & Baumann, 2020, which contains much more details on the concept of malevolence, the risks posed by malevolent individuals in power, and possible interventions to reduce the influence of malevolent actors.

[134] See Taylor, 2019.

[135] For more details and supporting references, see Althaus & Baumann, 2020, as well as Section 14.6 in Vinding, 2022a.

[136] See Caplan, 2006 for more details on this point.

this might enable a cruel individual in a position of power to create suffering on an unprecedented scale.

How risk factors interact

It would be misguided to view each risk factor as independent. Instead, there are numerous connections and complex interactions between the factors I outlined. For instance, polarisation and conflict can increase the likelihood that a malevolent individual rises to power. A dictatorship under a malevolent leader would, in turn, likely impede efforts to prevent s-risks. Advanced technology could potentially multiply the harm caused by malevolent individuals — and so on.

Conversely, the presence of a single risk factor can, at least to some extent, be mitigated by otherwise favourable circumstances. Advanced technological capabilities are much less worrisome if there are adequate efforts to mitigate s-risks. Likewise, without advanced technological capabilities or space colonisation, the suffering caused by a malevolent dictator would at least be limited to Earth.

It therefore seems plausible that most expected s-risks occur in worlds where several risk factors coincide. The risk might even scale in a superlinear way. This would mean that if two risk factors materialise, the likelihood of an s-risks is *more* than twice as high compared to a future where only a single risk factor materialises.[137]

[137] A similar pattern can be observed when it comes to risk factors in a medical context (which inspired this framework). A single medical risk factor (like age, obesity, or high blood pressure) is (in many cases) not yet catastrophic, but the combination of several risk factors often is.

CHAPTER EIGHT

Moral advocacy

Many factors can influence future outcomes. These include technological progress, economic dynamics, cooperation problems, and political or cultural trends. But perhaps the most fundamental determinant of how the future will go are the *values* of relevant decision-makers. We can only make progress on issues, from wild animal suffering to s-risks, if sufficiently many people *care* about them, thus creating the necessary political will to tackle these issues.

Advancing better values could therefore be a good lever for reducing s-risks. This relates directly to the risk factors for s-risks that we discussed in the previous chapter: better values increase the probability that adequate actions will be taken to prevent s-risks. In particular, better values make it more likely that advanced technology will be used responsibly, even if we cannot accurately predict how the world will change. We may be able to leave many future problems to future people — but only if they share our values.

Expanding the moral circle

Throughout the earlier chapters, we have seen that many s-risks relate to the disregard of the interests of non-human beings. Thus, a top candidate for what it means to promote "better values" is to ensure the moral consideration of *all* sentient beings. We should promote concern for suffering irrespective of who is experiencing it — no matter

who they are, what time they live in, or what species they belong to.

This approach has been termed *moral circle expansion*.[138] The "moral circle" refers to the set of beings or entities whom we grant moral consideration. An expanded moral circle would (so the argument goes) reduce the risk that neglected types of future sentient beings will be harmed on a large scale.

A key aspect of moral circle expansion is the rejection of *speciesism* — the discrimination against beings who belong to a different species.[139] As famously argued by Peter Singer in *Animal Liberation*, this can be viewed as a form of discrimination that is just as untenable as discrimination based on an individual's ethnicity, sex, or age.[140] Sentience should form the basis of moral consideration, not an individual's species membership (or any other morally irrelevant characteristic). We should give equal priority to equal interests, and to equal suffering in particular.[141]

Speciesist or anthropocentric attitudes are the key drivers behind our disregard of many nonhuman beings, as well as the horrendous harms we inflict on animals in factory farms and slaughterhouses. Thus, we can likely prevent a lot of future suffering if we manage to spread antispeciesist views and improve attitudes towards nonhuman animals.[142]

But the expansion of the moral circle goes further than that. Beyond the sentient beings inhabiting Earth today, we should also consider the possibility of novel forms of sentience that might emerge in the future. As discussed in Section 1.4, it is conceivable that sentient artificial entities could be created at some point. If that were to happen, it is crucial that we extend moral consideration to such artificial minds

---

138 Anthis & Paez, 2021.

139 For more details, see Vinding, 2015; Horta, 2022.

140 Singer, 1975.

141 See also Horta, 2010.

142 A key advantage of this approach, compared to advocacy for welfare reforms or a vegan diet, is that antispeciesism encompasses not only farmed animals, but also wild animals. Wild animals currently constitute the vast majority of sentient beings on Earth, and their suffering is particularly neglected (cf. Animal Ethics, 2020, and Section 11.2 in Vinding, 2020a). For these reasons (among others), Vinding, 2016b argues that animal advocates should focus more on the explicit promotion of antispeciesism.

(including disembodied or "voiceless" ones), especially if they are created in large numbers.[143]

That said, we must distinguish carefully between different aims. If our core aim is to reduce s-risks, we should be careful not to uncritically assume that expanding the moral circle is perfectly convergent with our core aim. And expanding the moral circle is, in turn, distinct from the goal of improving the welfare of farmed animals in the here and now. While these different goals are clearly related, it would be a remarkable coincidence if the same interventions happened to be ideal in terms of several distinct goals at the same time.[144] Instead, we will likely arrive at (somewhat) different interventions when focusing on s-risk reduction compared to if we were focusing chiefly on those other goals.

It would also be misguided, though, to conclude that animal advocacy efforts are unimportant from a long-term or s-risk-focused perspective. Most animal advocates currently focus on human-caused animal suffering in the (relative) short term, but their messages could be broadened to encompass all sentient beings, including wild animals and potential artificial minds. Likewise, establishing basic rights for some nonhuman animals could set legal and social precedents that may help us to later expand protections to other classes of sentient beings.

Risks of moral circle expansion

The basic argument for why moral circle expansion reduces s-risks

---

[143] For more details, see Harris & Anthis, 2021, as well as Section 11.4 in Vinding, 2020a.

[144] See Lewis, 2016 for more details on the idea of surprising or suspicious convergence.

On the other hand, it is less surprising if there is some degree of convergence on a broad category of actions for the near and long-term future. For instance, increasing moral consideration of neglected beings is probably a solid heuristic for improving the world, regardless of the timeframe. Also, if the current knowledge and resources of the animal movement are not yet strongly optimised, there is presumably more room for improvements to both short and long-term impact (e.g., increasing effectiveness in general).

seems intuitively plausible: a larger moral circle increases the likelihood that adequate measures will be taken to prevent s-risks — especially those that affect non-human beings who are often excluded from moral consideration.

Yet on closer inspection, the relationship between s-risks and moral circle expansion is not as straightforward as it might appear.[145] A larger moral circle could backfire in combination with the "wrong" values or beliefs. For example, concern for wild animals could, in combination with certain environmentalist values, result in increased efforts to preserve nature in its current state, in spite of the immense amounts of animal suffering in nature.[146] Or a larger moral circle could result in the creation of more (potentially suffering) sentient beings when combined with highly optimistic moral views according to which suffering can readily be outweighed by other goods.[147]

If not done carefully, moral advocacy also risks a backlash that could further entrench bad values or antagonistic dynamics. In the worst case, the idea of caring about all sentient beings might become a divisive "hot button" issue, akin to ongoing "culture wars". As discussed in the previous chapter, such polarisation would be a substantial risk factor for s-risks.

This can be a strong reason to avoid a confrontational approach. We can and should be assertive about our values, but we should also be friendly and cooperative in our efforts to convince others, so that we do not accidentally push people towards worse values. A grave deterioration of values could potentially result in agents with outright malevolent, vindictive, or sadistic attitudes and goals. And the attainment of the "optimal" values may be less important, in terms of s-risk prevention, than the avoidance of uniquely bad values.[148] This may be especially true if s-risks are heavy-tailed (as discussed in Chapter 5).

---

[145] Many of the following arguments are based on Vinding, 2020a, Section 11.6.

[146] See Tomasik, 2013b.

[147] See Vinding, 2018a.

[148] This point is also made in Vinding, 2020a, Section 9.6; Vinding, 2022a, Section 9.3.2.

Robust forms of moral circle expansion

The risks and caveats outlined above are reasons to doubt whether generic work to expand the moral circle is a top priority from an s-risk perspective. Still, expanding the moral circle seems likely to do more good than bad in expectation, and therefore remains an endeavour that is worth supporting.

Instead of refraining from moral circle expansion altogether, we need to find ways to expand the moral circle in thoughtful, sustainable, and prudent ways. Targeted forms of advocacy can be robustly positive for s-risk prevention if they avoid the drawbacks of moral circle expansion. Likewise, it can be robustly positive to improve and develop existing social movements, by increasing the degree to which they are able and motivated to reduce s-risks (in addition to other goals).

The animal advocacy movement seems particularly relevant in this context. In the following, I will outline suggestions for how the animal movement can become (even) more effective at reducing the suffering of all sentient beings, from a perspective that takes s-risks into account.[149]

First, it is vital that we as animal advocates always remain open-minded and willing to learn more. This relates not just to empirical facts or strategic considerations, but also to foundational philosophical beliefs. For example, many animal rights activists have given little consideration to issues such as the suffering of animals living in nature, the vast quantities of invertebrates such as insects,[150] or the possible sentience of artificial entities. If the animal advocacy movement is to be a movement for *all* sentient beings, we need to give deeper consideration to these neglected issues.

Second, the movement would do well to, on the margin, focus more on long-term considerations. While it is urgent and important to alleviate animal suffering in the short term, this should not be the sole focus of our efforts, given the potentially much larger number of future beings. A long-term outlook entails an emphasis on achieving lasting social change and on ensuring the long-term stability of the movement.

---

[149] For more details, see. Baumann, 2020c; Baumann, 2022a; Vinding, 2022a, Chapter 10.

[150] See Schukraft, 2019.

It is vital to avoid actions that impair our ability to achieve our long-term goals — as individuals, as organisations, and as a movement.

Third, it is probably best to avoid needless controversy and to advance concern for non-human beings in a non-partisan way. For example, it is helpful to emphasise that the cause of reducing animal suffering draws support from the entire political spectrum, including conservative and libertarian voices.[151] Such a cooperative approach reduces the risk of a serious backlash that could jeopardise our long-term influence. (To be clear, a cooperative approach is perfectly compatible with being assertive about the moral importance of all sentient beings.)

Research suggests that a main driver of the backlash and hostility of some meat eaters towards vegans and vegetarians is a perception of being judged as morally inferior.[152] So we should take care to avoid triggering such a perception — e.g., by primarily framing the issue of animal suffering in institutional or political terms, rather than in terms of individual food choices.

Promoting concern for suffering

Another promising strategy is the promotion of concern for suffering.[153] To counter the possible downsides of moral circle expansion, we could further develop and advance suffering-focused views (in addition to advancing moral consideration of all sentient beings).

A key advantage of this strategy is that people who endorse suffering-focused views are particularly likely to prioritise s-risk reduction. New information can always change our assessment of a concrete intervention or policy, but agreement at the level of fundamental ethical principles seems robustly good, even in the face of great uncertainty. So advancing suffering-focused views might be a

---

151 See Nozick, 1974, p. 38, and Scully, 2002.

152 Minson & Monin, 2012.

153 This section draws on Tomasik, 2015c and Chapter 12 in Vinding, 2020a.

uniquely robust and effective way to reduce s-risk.[154] (Of course, this strategy only makes sense if you endorse suffering-focused views yourself.)[155]

Promoting concern for suffering can take many forms. An important part is the refinement and dissemination of philosophical arguments that support a special priority to the reduction of suffering. To be clear, any exploration of suffering-focused ethics should always be done in an intellectually honest way, ideally with an emphasis on cooperation with other value systems.[156] (The Center for Reducing Suffering, which I co-founded, is an example of an organisation that follows this approach.)

Other options include raising awareness of real-world cases of extreme suffering, or helping people close the gap between their ideals and their actions — e.g., by overcoming defense mechanisms such as denial or wishful thinking.[157]

Similar to moral circle expansion, it is generally best to advance concern for suffering in a non-partisan and nuanced way, to improve the chances that suffering reduction becomes a common goal rather than a controversial issue.

Conclusion

Moral advocacy can be an effective strategy to reduce s-risks, especially if done in a careful and prudent way that avoids backfire risks.

Of course, this brief overview does not answer the question of whether moral advocacy is the *most* effective way to reduce s-risks or

---

[154] See Tomasik, 2015c for a more detailed argument, as well as a discussion of potential downsides.

[155] As discussed in Section 4.1, "Suffering-focused ethics", endorsing suffering-focused ethics is not necessary to favour work on s-risks. But a person who prioritises s-risks without endorsing suffering-focused ethics should, of course, not advocate a view that they do not believe in. (Instead, they could pursue other interventions to reduce s-risk.)

[156] See Tomasik, 2014.

[157] See Cohen, 2001; Tomasik, 2006.

improve the long-term future.[158] That depends on the tractability of long-term social change, the effectiveness of other interventions, the time-sensitivity of moral circle expansion, and many other factors.[159]

If we decide that promoting certain values is important, we still need to find the best ways to do so. For example, animal advocacy can take many forms, including corporate campaigns in favor of animal welfare reforms, efforts to advance research and development of plant-based or cultivated meat,[160] or philosophical arguments against speciesism. However, a detailed discussion of strategic questions relating to animal advocacy is beyond the scope of this book.[161]

---

158 For further details on this, see Chapter 11 in Vinding, 2020a and Chapter 10 in Vinding, 2022a.

159 For more details, see Baumann, 2017a.

160 Cultivated meat is meat produced from growing cells in a medium, without the slaughter of animals.

One might object that providing such alternatives to animal products only changes behaviour and will have no effect on our long-term attitudes towards animals. However, if we no longer need to rationalise eating animals, it will likely become psychologically easier for people to care about them. See Bastian et al., 2012.

161 See Anthis, 2017; Vinding, 2022a, Section 10.9.

CHAPTER NINE

Better politics

Our political institutions and discourse relate directly to several risk factors for s-risks. Dysfunctional politics can foster polarisation, thwart efforts to prevent s-risks, and increase the risk of malevolent actors rising to positions of power. Efforts to improve politics are therefore a plausible lever to tackle s-risks.

More broadly, a functional political system puts society in a better position to handle the challenges of the future (including the challenge of averting potential s-risks). Political decisions are the linchpin of our collective decision-making. And our political system is currently operating far from ideally, to put it mildly. We can and should do better.

How to achieve better politics is, of course, a complex question. It is impossible to cover all relevant aspects in this book, so I will focus on proposals that seem particularly promising or important from an s-risk perspective. For a more comprehensive analysis, I refer the reader to Magnus Vinding's freely available book *Reasoned Politics*.[162]

The two-step ideal

It is worth aspiring to a two-step ideal of politics that involves a

[162] Vinding, 2022a.

normative step followed by an empirical step.[163] The **normative step** is to clarify the aims and values that underlie our policymaking. This is not about merely stating our goals, but also about open-minded conversation and moral argument to discuss and refine the values that (should) form the bedrock of our collective decision-making.

Once we have identified a set of carefully reflected values, the **empirical step** is to ask which policies are optimal for achieving our aims. This is usually a complex factual question, which requires us to draw on the best available evidence and to engage in an open-ended scientific investigation and discussion.

In short, the two-step ideal recommends that we adopt the mindset of a moral philosopher (in the normative step) and then that of a scientist (in the empirical step). By contrast, contemporary political rhetoric often confuses empirical and normative aspects, which hampers clear thinking.

Following the two-step ideal would likely enable greater precision in our political conversations, as it helps us to clarify where we disagree and where we have common ground. This, in turn, allows for more fruitful political discourse, mutually beneficial compromises, and perhaps even moral progress.

Efforts to move closer to the two-step ideal go hand in hand with an awareness of the various biases that often prevent us from approaching policy questions with an open mind. The most common biases include confirmation bias (as discussed in Section 6.3) and motivated reasoning, which is when we seek to justify a desired conclusion rather than following the evidence where it leads.[164] Overconfident political views are also ubiquitous, in spite of the complexity of most policy questions and the fact that most voters are not well-informed about politics.[165]

This underscores the need for a deliberate effort to overcome biases and to approach political issues as open questions. By consciously seeking open moral argument and empirical evidence,

---

[163] This section closely follows Section 2 in Baumann, 2022b. For more details, see Chapter 1 in Vinding, 2022a.

[164] See Kunda, 1990.

[165] This is a robust finding in political science. See Bartels, 1996 and Gilens, 2001. Political ignorance also affects big issues, not just minor details.

rather than relying too strongly on immediate intuitions, we can likely reduce political overconfidence. In particular, it seems beneficial to focus more on understanding *how* a suggested policy works, rather than jumping to support or oppose it.[166]

What does an improved culture of political discourse have to do with s-risk reduction? It is, of course, not easy to pinpoint or quantify the effects. Yet it seems plausible that more reasoned political thinking and debate would serve as an antidote to the risk factor of (excessive) polarisation, and that it reduces the likelihood that populist demagogues with malevolent traits are able to rise to power. Greater adherence to the two-step ideal would therefore likely reduce s-risks — while also contributing to better outcomes more broadly.[167]

Overcoming our tribal nature

Muddled thinking or a lack of information are arguably not the only distorting factors in contemporary politics. Another core problem is our tribal nature.

A key finding of modern political science is that social attachments to groups are among the most important factors determining our political judgments.[168] This is in line with the idea that the primary purpose of our political behaviour is *loyalty signalling*: to express our allegiance to a party, social movement, or other community.[169] According to this theory, political debate is not chiefly about promoting good policies or arriving at some notion of truth. Instead,

---

[166] Fernbach et al, 2013, find that extreme or overconfident attitudes are often based on an illusion of understanding. When people are asked to explain how a policy works — rather than just give reasons to support or oppose it — their self-rated level of understanding drops, and people's views become significantly more moderate. This suggests that a greater focus on how a policy works could help to reduce political overconfidence. See also Vinding, 2022a, Section 4.6.

[167] I also recommend Freinacht, 2017 and Freinacht, 2019 for an in-depth philosophy of how society and political culture could be developed further.

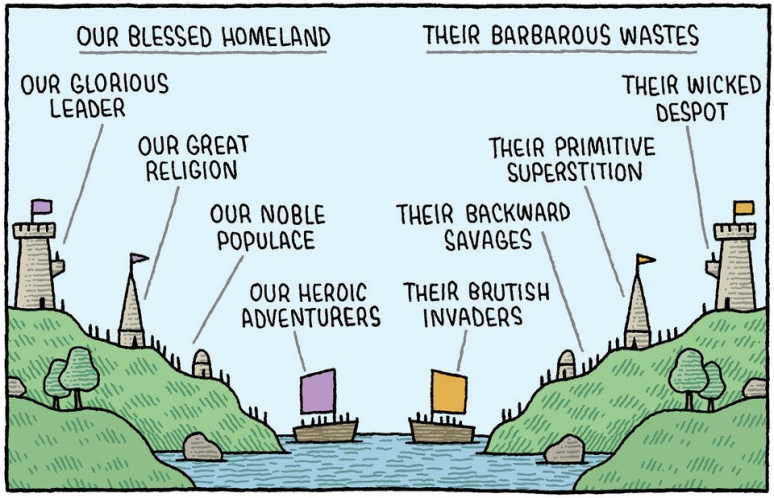
[168] Achen & Bartels, 2016.

[169] For more details, see Simler & Hanson, 2017, Chapter 16; and Vinding, 2022a, Chapters 2-4.

we mostly cheer for our team and boo their team in a zero-sum game.[170]

A closely related concept is *hot cognition*: our brains tend to automatically process political individuals and issues in emotionally charged ways.[171] We instinctively view our own political tribe and leaders in a positive light and "the other side" in a negative light. In the worst case, this can spiral into the vilification or scapegoating of certain individuals or outgroups (e.g. ethnic minorities).

Our propensity for loyalty signalling and hot cognition explains many of our biases. For example, the drive to signal loyalty often leads us to display a high degree of (over)confidence in the core tenets of our ingroup and hostility towards rival groups. Expressions of uncertainty and nuance do not fit this team sports mentality — every argument, and even the "facts", must favour one's own team, lest one is (seen as) disloyal.



(Cartoon by Tom Gauld, 2015.)

Considering how ubiquitous and detrimental these dynamics are, it is imperative that we resist the pull of loyalty signalling.[172] This will

[170] This does not require us to be aware of these motives. Simler & Hanson, 2017 argue that we are often self-deceived about what animates our behaviour, both in political and in non-political contexts.

[171] Lodge & Taber, 2005.

[172] See also Vinding, 2022a, Section 4.3.

be challenging, as it requires us to overcome tendencies that are deeply ingrained in our psychology.

A helpful strategy is a heuristic in favour of nuance.[173] If we acknowledge grains of truth in different perspectives, and think in terms of degrees of credence rather than rigid certainties, we can avoid us-versus-them and black-or-white thinking.

Another good heuristic is a norm of being charitable and respectful toward political opponents, by engaging with the strongest interpretation of their views and arguments. And again, it is worth emphasising the shared goal of better understanding *how* a given policy works, instead of rushing to divide people into supporting and opposing camps.

Overcoming our tribal instincts is directly relevant to s-risk reduction. Limiting the influence of biased intuitions and dogmatic partisan loyalties helps avoid excessive political polarisation (which constitutes a risk factor for s-risks, as discussed in Chapter 7). And more charitable political conversations arguably make escalating zero-sum conflicts and resulting worst-case outcomes less likely. Raising the standards of political discourse could therefore be a promising way to reduce s-risk.

Reducing tribalism also opens up avenues for cooperation and compromise, which increases the likelihood that everyone's moral concerns are taken into account.[174] To be clear, a certain degree of political competition is both healthy and unavoidable. Yet politics can often be win-win if we think in terms of policy outcomes and not in terms of "beating" the other faction.

It is worth noting that extensive loyalty signalling frequently creates the *appearance* of major disagreements when there are actually only limited disagreements in terms of policy substance.[175] We are often less divided than we think. Indeed, the prevention of (severe) suffering is a good example of a widely shared and uncontroversial value.[176]

---

[173] See Vinding, 2018b.

[174] Vinding, 2020b argues that cooperation is particularly important from an altruistic point of view.

[175] Hannon, 2021.

[176] See Vinding, 2022a, Chapter 7 for more details.

This is also a reason to focus political discourse more on policies and less on parties or individuals. The latter lends itself to tribal mudslinging, smears, and name-calling, while discussions of policy substance tend to be more fruitful — especially when we focus on mutually beneficial policies.

Institutional reform

The previous sections focused on our political culture. Yet our political institutions are perhaps just as vital. For example, fundamental principles such as democracy, the rule of law, and basic human rights are arguably essential to achieving stability, prosperity, and liberty. To see how important these principles and the corresponding instutions are, one only needs to compare North Korea and South Korea.

From an s-risk perspective, important goals are to prevent excessive polarisation, to keep malevolent individuals in check, and to prevent a descent into totalitarianism. Functioning democratic institutions can provide checks and balances to ensure that any single individual can never gain too much power, thus reducing the influence of malevolent actors. Conversely, a dysfunctional system often allows the most ruthless and strategic individuals (with malevolent traits) to rise to power, resulting in a "pathocracy".[177]

Efforts to strengthen democratic governance can therefore be a promising lever for reducing s-risks. Of course, myriad proposals for how to improve governance have been brought forward, and a detailed analysis is beyond the scope of this book. In the following, I will focus on a brief overview of reforms that are particularly evidence-based and relevant for s-risks.[178]

---

[177] The term pathocracy means rule of psychopaths — or more precisely, individuals with malevolent personality traits. See Taylor, 2019, as well as Section 14.6 in Vinding, 2022a.

[178] For a more extensive analysis, I refer the reader to Chapter 14 in Vinding, 2022a.

Parliamentarism

An example of a promising institutional reform is the switch to a parliamentary system of government in countries that currently use the presidential system.[179] A growing body of evidence suggests that the parliamentary system is superior in many ways. In particular, research indicates that countries using the parliamentary system exhibit lower levels of political polarisation, less economic inequality, and a lower risk of democratic backsliding compared to presidential countries.[180]

The parliamentary system is also likely better in terms of keeping malevolent actors in check.[181] This is because power is more decentralised and the head of government can be dismissed fairly easily, whereas the President is usually elected for a fixed term in a presidential system.

Overall, it seems plausible that parliamentarism mitigates several key risk factors for s-risk. But there is still significant uncertainty about the effect size, the best ways to promote parliamentarism, and the tractability of switching systems in countries that have a long presidential tradition (like the US).

Voting reform

Another institutional issue in modern democracies is the use of majoritarian voting systems. An example is the plurality voting or "First past the post" system, which determines a single winner per district. Everyone gets a vote, and the candidate with the most votes wins. This system is used in many major democracies, including the United Kingdom, the United States, and India.

---

[179] In a parliamentary system, the executive (in particular, the head of government) is accountable to the legislature (i.e. appointed and potentially dismissed by majority vote in parliament). By contrast, in a presidential system, the head of government (the President) is elected directly by the people and not accountable to parliament.

[180] This is argued in detail in dos Santos, 2020. See also Section 14.5.3 in Vinding, 2022a.

[181] However, this has (to my knowledge) not yet been researched, in part because it is hard to reliably diagnose malevolent traits. So this point is more theoretical and less evidence-based than the previously mentioned points.

An advantage of plurality voting is its simplicity. However, plurality voting usually does not achieve representative results, as the share of seats won by a party can diverge significantly from its vote share. According to Duverger's law, the winner-takes-all nature of the system tends to create a two-party system.[182] This can contribute to excessive polarisation because a two-party system feeds into our tribal nature. If there are only two opposite poles, political actors have an incentive to focus more on demonising the other party than on achieving anything of substance.[183] A prominent example of this dynamic is the excessive polarisation between Democrats and Republicans in the United States.

By contrast, proportional voting systems (which translate the share of votes into a roughly equal share of seats) usually result in a multi-party system. If no party achieves a majority on their own, parties have to work together and compromise. This does not always work perfectly, but the shifting coalitions still reflect a more cooperative model of politics and mitigate the tribal "us-versus-them" mentality.

This story is borne out by the available evidence. Countries that use proportional representation generally feature less political polarisation, higher voter turnout, higher satisfaction with democratic institutions, less economic inequality, along with various other advantages.[184] Research also suggests that democracies that use a proportional voting system have significantly lower war involvement compared to democracies that use a majoritarian system.[185]

Proportional representation is closely linked to parliamentarism. A presidential election necessarily has a single winner and is therefore, in a sense, maximally disproportional. So the advantages of proportional representation cannot be fully realised in a presidential system. This suggests that the ideal endpoint of reform is a proportional parliamentary system.[186]

---

182 Duverger, 1954.

183 See Drutman, 2020 for more details.

184 See Fairvote Canada, 2018 for an overview.

185 Leblang and Chan, 2003.

186 Vinding, 2022a, Section 14.5.4 reaches a similar conclusion. It is worth noting, though, that proportional parliamentarism is just taken to be the ideal

It is worth noting that ongoing voting reform efforts often attempt to replace plurality voting while retaining a (non-proportional) system based on single-winner districts. For instance, many groups in the United States focus on the introduction of ranked choice voting.[187] While ranked choice voting would likely be a modest improvement over the status quo of plurality voting, it would likely fail to fully bring about the above-mentioned benefits of a proportional system.[188] So it seems preferable to focus on more ambitious reforms.

Advancing and safeguarding democracy

From an s-risk perspective, it is perhaps more important to avoid catastrophic institutional failures than to strive for the best possible institutions. The difference (in s-risk terms) between a democracy and an autocratic or totalitarian system may be larger than the difference between a functional democracy and a flawed democracy.[189]

Modern liberal democracies offer a much better protection of human rights and civil liberties like free speech. These fundamental rights are a precondition for being able to raise moral concerns, which is one reason why their frequent suppression in autocratic systems constitutes a risk factor for s-risks.[190]

Democratic principles and institutions have also been designed to have checks in place to reduce the influence of malevolent individuals. While democracies are not always successful in preventing malevolent

---

within the hitherto explored space of democratic institutions. A further question is whether more radically different institutions, such as randomly selected citizen's assemblies or a complete abolition of the state (i.e., anarchy), could be better still. Yet this is more speculative and we lack empirical data on any such alternative system. See Vinding, 2022a, Section 14.4 for more details.

[187] An election that uses ranked choice voting (also called instant runoff voting) allows votes to rank candidates. The votes of eliminated candidates are then transferred to the second (or third, etc.) preference, until a candidate wins an outright majority.

[188] Drutman & Strano, 2021 review the evidence on ranked choice voting and argue that the benefits are relatively modest.

[189] It is worth noting, though, that institutional flaws and resulting issues like excessive polarisation also increase the likelihood that a democracy fails and is replaced by an authoritarian system.

[190] This is argued in Vinding, 2022a, Chapter 11 and Section 14.3.

rule, they appear better than any other system in this regard. So it seems likely that democracy helps reduce s-risks — taking into account, also, that war between democratic countries seems uniquely rare.[191]

This leaves open how we can best promote democracy.[192] It is unclear whether we should focus most strongly on advancing liberal democracy in currently non-democratic nations, on safeguarding existing democracies against democratic backsliding, or on strengthening democracy in semi-democratic states. In addition, it is critical to ensure that promoting democracy is not used as a pretext to pursue economic or foreign policy interests. Such misuse is highly unfortunate as it can discredit the idea of advancing democracy, and even contribute to great power conflict.

Political representation of all sentient beings

Another proposal is the idea of extending political representation to all sentient beings, including non-human animals. Non-human animals cannot represent their interests themselves, but we could, for instance, institute commissioners on their behalf who are solely tasked with defending the interests of non-human animals. This has been termed "sentientist democracy" or "sentiocracy".[193]

Similarly, an argument can be made that we should grant representation to future individuals — another class of individuals that go wholly unrepresented in current institutions.[194]

The political representation of all sentient beings would likely reduce s-risks because it would — similar to moral advocacy — result in better protections of previously neglected individuals, at least in

---

[191] There is an ongoing academic debate on "democratic peace theory", particularly regarding the causality of the association. See Imai & Lo, 2021; also Vinding, 2022a, Section 14.3.4.

[192] See Schonfeld, 2020, as well as Vinding, 2022a, Section 14.7.2.

[193] See Donaldson & Kymlicka, 2011; Cochrane, 2018; Vinding, 2022a, Section 14.5.1.

[194] However, if the representation of future generations remains limited to future humans, while excluding nonhuman animals, then this proposal might not be very beneficial from an s-risk perspective. See Baumann, 2020d for more details.

theory. On the other hand, a "sentiocracy" would be relatively uncharted territory compared to the other reform proposals listed above, which have already proven successful in many countries and are backed by tangible evidence.

Is politics too crowded?

I have so far bracketed questions about the effectiveness or tractability of efforts to improve politics. One might argue that the area is too crowded due to a plethora of actors vying for political influence. And if there were any easy solutions to political dysfunctionality, why have they not already been found and implemented?

There is an element of truth to this, but I believe that efforts to improve politics can still be valuable for many reasons.

While many people pursue some political agenda, a much smaller number is systematically working to improve our political culture or institutions in an effective and evidence-based way, drawing on the best available scientific insights. And aspects that relate to s-risk in particular — like malevolent traits, or the representation of non-human beings — are more neglected still.

Also, even a marginal improvement of political culture (or institutions) can potentially be highly beneficial. Transformative change will not happen overnight, but it is quite realistic that we could become *somewhat less* biased and *somewhat less* tribal, or that our institutions could be *somewhat more* functional. We can plausibly increase the *degree* to which our politics is based on carefully reflected values and sound empirical evidence.

Conclusion

While it is difficult to know what the ideal political institutions or culture would look like, the bar for doing better than the status quo seems low. And another reason for optimism is that many insights about our political psychology and biases are fairly recent and have not yet fully trickled into the public's awareness. (Likewise, evidence-based findings on the benefits of parliamentarism or proportional representation are yet to reach a wider audience.)

A key advantage of many of the above proposals is that they are beneficial from many perspectives and not only in terms of reducing s-risks. This chimes with a cooperative approach to s-risk reduction and increases the likelihood of gaining enough traction to achieve positive change.

Finally, it is often important to state and aim for ideals even if their large-scale adoption is not feasible in the foreseeable future. Continuous modest improvements can, over a sufficiently long time horizon, add up to transformative change. And a good starting point is to follow the two-step ideal in our own thinking and communication, and to set an example by advancing the principles of reasoned political discourse within our own communities.[195]

[195] See also Vinding, 2022a, Section 1.7.

CHAPTER TEN

Emerging technologies

I have so far not addressed the risk factor of advanced technology. If transformative technologies emerge in the future, it seems plausible that shaping their development could be a good lever to influence the long-term future and to reduce s-risks.

I will bracket attempts to stop technological progress altogether. This may, in theory, help prevent s-risks, but it is unrealistic to expect that we could stop all progress forever (even if it may be feasible to prevent certain kinds of harmful technologies). Moreover, trying to stop progress is likely to provoke conflict with those who want to reap the benefits of advanced technology.[196]

Instead, my focus will be on measures to ensure the prudent and responsible use of potentially pivotal technologies.

Should we focus on shaping artificial intelligence?

Advanced artificial intelligence (AI) is one of the most frequently

[196] It is worth keeping in mind that technology can also be used to reduce suffering (in addition to being a risk factor for s-risks, as discussed in Chapter 7). An example is cultivated meat, which has the potential to render conventional animal farming obsolete. More advanced technology may also facilitate interventions to reduce the suffering of animals in nature. See Johannsen, 2017.

discussed examples of a potentially transformative technology.[197] If we create systems that can surpass humans in terms of general intelligence, then this would (so the argument goes) constitute a pivotal event in human history. Such a system could, due to its superhuman intelligence, effectively take over the world and shape everything that happens later on according to its values.

If it is plausible that we create extremely powerful AI in the not-too-distant future (say, this century), it stands to reason that influencing this process is a top priority. Shaping the development and use of advanced AI could be an extraordinarily consequential lever to influence the long-term future.[198]

Indeed, efforts to reduce s-risks caused by AI may be considered more targeted or more direct than the other interventions I have discussed so far, such as moral advocacy or better politics. After all, a large share (perhaps a majority) of expected future suffering stems from scenarios that feature advanced technology (cf. Chapter 7).[199]

Work to avert s-risks caused by AI is also highly neglected. A small number of people do research on AI safety, a field that centers on the question of how to ensure that the goals of autonomous AI systems are aligned with those of its creators.[200] Yet alignment would likely not be sufficient to prevent potential s-risks resulting from an imprudent development or use of powerful AI, which suggests that more targeted measures are needed.[201] Since few people have, so far, considered AI safety from an s-risk perspective, we can expect that there are still low-hanging fruits left to reap.

---

[197] Other candidates for transformative technology include embryo selection for cognitive enhancement (Shulman & Bostrom, 2014) or whole brain emulation (cf. Hanson, 2016), although the latter arguably constitutes a form of artificial intelligence. For simplicity, I will limit the discussion to advanced AI.

[198] For a more detailed argument, see Gloor, 2016a — as well as Vinding, 2018c for a critical reply.

[199] It is worth noting that this might hold even if the probability of such scenarios is low.

[200] This has been termed the Alignment Problem, cf. Christian, 2020.

[201] Even if an aligned AI is, all things considered, less likely to lead to s-risks, it would be a striking coincidence if alignment work is also most effective for s-risk reduction. That said, AI alignment and s-risk-focused AI safety are generally complementary. See also Sotala & Gloor, 2017.

On the other hand, there are good reasons to be sceptical about predictions regarding future technology, especially when the claims in question are far-reaching. Considering the current capabilities of artificial intelligence and the pace of progress in the field, a rapid takeover of human civilisation by superintelligent AI seems unlikely to me, at least in the foreseeable future.[202]

In general, we do not know much about what future technologies will look like. Details of technological advances are hard to predict, and we lack clear feedback loops or an empirical grounding to make confident predictions. So the risk is high that our efforts to directly shape AI will be ineffective, if not wasted altogether, as we have limited knowledge about the nature of this technology.[203] Or we might be misguided about which future technologies will be most consequential.

We also need to unpack concepts like "intelligence" (and, by extension, "superintelligence"), as they are often used in a broad and simplistic way. For instance, one might think that the capabilities of modern civilisation are exclusively owed to the great power of individual human intelligence. But a closer look reveals that individual intelligence is just part of the picture, and that social learning and cultural accumulation of more and more tools are, in large part, the secrets to our success.[204]

It is also doubtful whether advanced AI will be a single, unified, all-purpose system. The trends we have seen so far in automation and software lend more support to a progressive increase, through many distributed innovations, in the *collective* capabilities of our civilisation, rather than a sudden jump in the capabilities of a single actor.[205]

AI will likely become superhuman in one domain after another (as

---

[202] To be sure, I believe that AI progress will likely continue. After all, it would be surprising if progress in the field stopped altogether. But that is a much weaker claim.

[203] Similar points are made in Vinding, 2020a, Section 12.4; Hanson, 2022.

[204] For a detailed defense of this claim, see Henrich, 2015. The implications in terms of the future of AI are explored in Vinding, 2016a.

[205] Arguments against a sudden jump in capabilities can be found in Vinding, 2016a and Vinding, 2018c. By contrast, Sotala, 2017 argues that the rapid development of artificial superintelligence is feasible. See also Hanson & Yudkowsky, 2013 for a comprehensive discussion.

it has in various domains for decades), rather than suddenly becoming superhuman in all domains at the same time. In this case, the emergence of powerful AI would not be an isolated event — instead, it would happen gradually. This would likely mean that there is no single exceptional lever or "silver bullet" to shape the future and reduce s-risks.[206] (Likewise, there was arguably no single leverage point or pivotal technology in the past that determined the course of history.)

Another key consideration is that technological developments are never divorced from their political and sociocultural context. Efforts to improve values or political culture (as opposed to directly shaping AI) can therefore also indirectly influence the emergence of powerful AI. Indeed, even if there is just a single leverage point, and even if we know what it will look like, it is still possible that improving background factors like our values or institutions might be the most effective way to shape the outcome.[207]

Where does this leave us? Despite the objections outlined above, it is still plausible that AI will be a critical future technology. AI could "take over the world" in the same sense in which the computer or the internet "took over the world" — gradually and over time. And given the far-reaching consequences of such a transformative technology, research into mitigating risks (as well as better understanding the nature of the risks) might be worthwhile despite considerable uncertainty.

Technical measures to reduce s-risks from AI

What could an s-risk-focused approach to AI safety look like?[208] I have mentioned before that destructive interactions and escalating conflicts can pose a serious s-risk (in Chapter 2). This risk seems

[206] Vinding, 2020a, Section 9.4 argues that we need to resist the temptation to seek a single, simple answer to the question of how to reduce suffering, and calls this hope the "silver bullet delusion".

[207] See Vinding, 2022d for additional remarks on what does and does not follow from the premise that AI will play a pivotal role in the future.

[208] This has been termed suffering-focused AI safety or worst-case AI safety. See Gloor, 2016c; Baumann, 2018b.

particularly pronounced in interactions between advanced AI systems, as they constitute a new and powerful type of actor.

Interventions to achieve *cooperative AI* therefore seem particularly promising for s-risk reduction. The goal of research on cooperative AI is to not just build intelligent systems, but to equip these systems with the necessary techniques and methods to achieve mutually beneficial outcomes in interactions with other agents.[209] For example, research on bargaining, game theory, or decision theory (in an AI context) could yield insights that help prevent negative-sum dynamics and navigate cooperation problems more productively. Likewise, a careful design of the training environment can help mitigate the risk that some analogue of malevolent traits might evolve in AI systems. These lines of research, which aim to avoid s-risks resulting from failures of cooperation or escalating conflicts among powerful AI, are pursued by the Center on Long-Term Risk.[210]

A concrete example of a promising safety measure is the implementation of a *surrogate goal* in AI systems. The idea is to add to one's current goals an additional surrogate goal that one did not initially care about. In case of escalating conflicts, adversarial actors will (in theory) target this surrogate goal rather than what one initially cared about. This could (subject to further research) help limit the downsides of cooperation failures by deflecting the disvalue resulting from adversarial dynamics onto the surrogate goal.[211]

Another intervention is to prevent the emergence of malevolent heuristics or "character traits" of newly trained AI systems. Just like dark tetrad traits evolved in human evolution, multi-agent AI training environments could incentivize alien equivalents of these dark tetrad traits in the motivations of AI systems. By studying the incentivizes reinforced in the training environment of transformative AI systems, one could try to make sure that such anti-social tendencies don't show

---

[209] See Dafoe et al., 2020; Dafoe et al., 2021; Baumann et al., 2020.

[210] Clifton, 2020 describes potential research avenues on cooperation, conflict, and transformative AI in more detail.

[211] See Baumann, 2017b for more details on various issues we face when attempting to specify and implement surrogate goals. Baumann, 2018a also discusses open questions and possible directions for further research. Oesterheld & Conitzer, 2022 introduce the framework of *safe Pareto improvements*, which can be seen as a generalisation of surrogate goals.

up in the first place.

We can also take inspiration from common engineering approaches to improve the reliability of safety-critical systems, such as *redundant* and *fail-safe* designs. Redundant designs duplicate important components of a safety-critical system to serve as backups in case of primary component failures (e.g., backup power generators), while fail-safety refers to design features whose sole function is to limit the extent of harm in case of a particular type of failure (e.g., fire suppression systems). By analogy, we might attempt to specifically limit the extent of damage in the event of a failure to align advanced AI with human values. And we could attempt to make AI safer by combining many different safety measures that kick in if the system fails.[212]

Yet a complete and watertight solution to the perils of advanced AI remains elusive. Many, if not all, of the suggested technical interventions may turn out to be infeasible. And it is possible, in light of great uncertainty, that safety efforts could backfire and inadvertently increase the risk of escalating conflict. So we need further research into how to best prevent s-risks caused by AI.

Governance of AI

Technical safety measures are only one side of the coin. It seems just as important to advance good governance mechanisms that help navigate the transition to a world with advanced AI. In particular, we should establish norms, policies, and institutions that help prevent worst-case risks due to AI.

For example, we could promote international cooperation to prevent escalating arms race dynamics between competing nations. This could be achieved through agreements (akin to other international treaties) to ensure the prudent and cooperative, rather than hasty and adversarial, development of advanced AI. International bodies could be established to oversee AI development.[213] Likewise, AI companies could agree on formal or informal rules about how to proceed if an

[212] For more details on possible safety measures, see Gloor, 2016; Baumann, 2018c.

[213] Erdélyi & Goldsmith, 2022.

artificial intelligence approaches certain capabilities. (Of course, any such agreement might be difficult to enforce.)

It is worth noting that s-risks from AI would likely take place in circumstances that differ radically from our current world. The emergence of a transformative technology (such as advanced AI) carries a significant risk of political turmoil, which means that things we normally take for granted, like the rule of law or democratic institutions, might break down. This, in turn, is a risk factor for s-risks. So it is valuable to look into how political turmoil in a potential transition period to advanced AI can be avoided, and how we can ensure that laws, rules, or agreements will still apply in this context.

Another aspect that seems particularly important from an s-risk perspective is the potential misuse of AI by malevolent actors.[214] As part of efforts to reduce the influence of malevolent individuals, it is vital to limit their ability to misuse advanced technology.[215] For instance, it might be worthwhile to raise awareness of the risks of malevolent individuals within AI companies, or to establish protocols to make AI-related institutions less susceptible to adversarial actors. And we could try to distribute future technological capabilities in a way that makes it difficult or impossible for any single actor to cause a lot of harm (including, but not limited to, s-risks).

Finally, we could try to expand the moral circle of the creators of AI — e.g. by establishing standards for how to proceed if an artificial intelligence appeared to demonstrate some degree of sentience.[216] This connects a focus on advanced AI with moral advocacy. However, efforts to promote the moral consideration of artificial sentience must be planned carefully, lest we inadvertently cause a backlash (as discussed in Chapter 8). It seems most robust to focus on further

---

[214] For more details, see Brundage et al., 2018.

[215] It is also conceivable (albeit speculative) that an AI system would develop malevolent traits, in so far as that concept can meaningfully be applied to nonhuman intelligences. This risk could be mitigated through careful monitoring of the training data and environment.

[216] A concrete proposal is a moratorium that would prohibit research that aims at (or knowingly risks) the emergence of artificial minds. Such a moratorium could be lifted if and when artificial sentience is better understood, and measures are put in place to avoid potential suffering. See Metzinger, 2021.

research and thereby contribute to a nuanced and reflective discussion of issues relating to artificial sentience.

Space governance

So far, I have focused exclusively on risks from artificial intelligence. But the risk factor of advanced technology also entails other aspects. In particular, many s-risks are linked to a large-scale colonisation of outer space, which suggests that *space governance* could be another potential focus area for s-risk reduction.[217] Space governance encompasses the laws, rules, norms, and institutions that structure interactions in space, as well as mechanisms that are used to establish and enforce those.

We currently lack a global framework for space governance. As of now, space is mostly a free-for-all, which poses a risk of race dynamics and severe conflicts. So it could be valuable to replace the current state of ambiguity with coherent regulations that contribute to positive (long-term) outcomes if and when large-scale space colonisation becomes feasible. To this end, we can do research to find out which governance mechanisms are most suitable, and then lobby for corresponding treaties or conventions.

One could also think that the easiest way to avoid s-risks is to oppose space colonisation altogether. Yet it is important to be pragmatic. If space colonisation is inevitable, then dogmatic opposition may be counterproductive, especially if it antagonises those who wish to colonise space. So it seems more fruitful to emphasise the idea that we should only embark on such an astronomical endeavour after having done everything we can to ensure a positive outcome.

Put simply, we should get our house in order on Earth before we consider spreading into space.[218]

---

[217] See Baumann, 2020e.

[218] For more details on ethical and strategic considerations relating to space colonisation, see Tomasik, 2013c; Vinding, 2020a, Section 14.2 and Section 14.3.

CHAPTER ELEVEN

Long-term impact

We tend to mostly think about what we can do right now. Yet it pays to adopt a long-term perspective in our efforts to tackle s-risks, as the best opportunities might arise in the (distant) future. Indeed, it would be a striking coincidence if, among all possible times, today is the single most effective time for efforts to avert s-risks.[219]

Capacity building

The most important thing to do now might be to ensure that future actors will be both motivated and able to prevent s-risks. Through such *capacity building*, we will be prepared to make the best possible use of opportunities to reduce s-risks when they arise in the future. So perhaps we should now invest in flexible capacity to multiply our future impact.

This is a uniquely robust strategy to deal with our vast uncertainty about the future. After all, we know relatively little about what the future will look like, about how we can best influence it, and about how we can best reduce s-risks in particular. This is especially true for s-risks that lie in the distant future or involve advanced technology. Yet we can still, despite great uncertainty, have a reliably positive impact

[219] See MacAskill, 2020 for an analysis of whether we are living in the most influential period in history.

by accruing flexible resources until we see more clearly how worst-case outcomes might occur and what potential countermeasures may be available.[220]

Capacity building can take many forms. For example, money is a flexible resource that can be transferred to the future through saving and investing.[221] If long-term financial returns outpace inflation, we can potentially accumulate a large sum of money to spend on s-risk reduction at a later point.[222]

But capacity building is about much more than purely financial investments. In fact, many of the interventions I have discussed so far also help build capacity. If we promote greater concern for suffering and the moral consideration of all sentient beings, we increase the motivation of future people to prevent s-risks that affect nonhuman beings. And if we improve our political culture and institutions, we improve the capacity of future civilisation to address any future challenge, including potential s-risks.

A movement to reduce s-risks

A key dimension of capacity building is movement building: to foster a community of people who are interested in and knowledgeable about s-risk reduction. After all, the current degree of

[220] Similarly, Vinding, 2022e argues that radical uncertainty about outcomes need not imply an absence of robust strategies to reduce suffering (e.g. capacity building).

[221] The approach of investing and compounding money over long timescales has been termed *patient philanthropy*. See Trammell, 2021.

[222] This strategy of investing over the long term has actually been implemented successfully by Benjamin Franklin. He made bequests of the equivalent of $100,000 each to the cities of Boston and Philadelphia, specifying that the funds were to be invested for 200 years. After that period, the compounded value had grown to $4.5 million in Boston and $2 million in Philadelphia. See Schwartz, 2022.

However, the strategy also comes with significant risks. For instance, investments might become worthless in the event of a financial crash or political turmoil. Overall, it seems unclear whether we should invest more or spend more now (relative to the status quo).

concern for s-risks seems clearly insufficient. By making reasoned arguments for why s-risks should be taken seriously, we can connect with a broader set of people.[223] Ideally, we can build a growing movement that will help steer humanity away from a future moral catastrophe.

Beyond growth, a main priority is to ensure the long-term stability of the movement to reduce s-risks. We should be mindful of the possibility that ideas relating to s-risks might evolve in undesirable directions. In the worst case, the movement could fall into disrepute, which could result in a lack of efforts to prevent worst-case outcomes. (This is similar to concerns about how moral advocacy might engender a backlash, as discussed in Chapter 8.)

To reduce this risk, it is important to establish healthy social norms among those who seek to reduce s-risks. A friendly and welcoming community is more likely to draw people in. And a cooperative attitude towards those who pursue other priorities tends to be more productive than a provocative or antagonistic approach. After all, virtually everyone can agree that s-risks should be prevented, so it is often possible to find common ground.[224]

Finally, it is vital that the movement to reduce s-risks is *cause-neutral*. In light of great uncertainty, we should not let our priorities be dictated by personal attachments to a particular cause. Akin to individuals who may be biased in their judgments (cf. Chapter 6), an entire movement can also focus prematurely on a narrow cause. To achieve its full potential, the movement must be able to update its views and change course based on new evidence. We need to react flexibly to social and technological developments to find the most effective ways to prevent s-risks.

---

[223] That said, it would be misguided to think purely in terms of convincing the largest number of people. What matters most is arguably to reach those in positions of influence. For instance, in scenarios that involve advanced AI, it seems most important to raise concerns relating to s-risks among relevant stakeholders and decision-makers.

[224] See Baumann, 2020a; Vinding, 2020b.

Research on how to best reduce s-risks

Movement building primarily addresses the *motivation* to avoid a future moral catastrophe. Yet our *competence* in this endeavour is just as important. Expanding our knowledge is a key dimension of capacity building, as it puts compassionate future actors in a better position to reduce s-risks.[225] (One could call it "wisdom-building".)

Given our great uncertainty about the future, any attempt to reduce s-risk requires careful thought. Indeed, the priority areas outlined so far (in Chapters 8-10) are merely current best guesses. It remains unclear which of these areas are most promising — and perhaps some other, unknown intervention is best. Needless to say, it would be a mistake to invest most of our resources on a cause that further research would reveal to be misguided. This highlights the need to better understand how our actions affect the probability of a future moral catastrophe.

So perhaps the most important thing to do is to figure out what to do. We need an open-ended research programme on how to best reduce s-risks, drawing on a variety of disciplines ranging from computer science to the social sciences. A top priority is to find out which s-risks are most likely, most tractable, most neglected, or of the greatest magnitude. This information will, in turn, feed into an analysis of the best practical ways to mitigate s-risks. After all, the end goal is to make a difference in the real world, so it is vital that we bridge the gap between abstract philosophical ideas and concrete interventions.[226]

That said, research is not a silver bullet. In light of the complexity of the question of how to best reduce s-risks, it can be hard to make substantial progress. And the challenge is compounded by the fact that researchers need to take strange future scenarios seriously while not getting bogged down in overly speculative ideas. This is a difficult balance to strike.

Nevertheless, my tentative conclusion is that learning more about how to best reduce s-risks should be one of our top priorities at this point. Research on this question is at an early stage, so we can

[225] See also Vinding, 2022a, Section 9.3.3.

[226] For an overview of open research questions, see Center for Reducing Suffering, 2021.

probably still find key insights that greatly increase the effectiveness of efforts to reduce s-risks. And even if groundbreaking insights elude us, research and reflection still help us to refine our views and to arrive at more nuanced and balanced conclusions.[227]

A strategy that focuses on research or other forms of capacity building may seem highly abstract. Such activities are unlikely to pull on our heartstrings. But considering the potentially astronomical stakes and the counterintuitive nature of s-risks, we cannot afford to blindly follow our most immediate emotions. We need to think carefully about what is most effective.[228]

What you can do

Are you interested in getting involved in reducing s-risks?

A simple first step is to join the discussion and get in touch with other people who are interested in reducing s-risks.[229] Being part of a community enables connections with potential mentors and can boost one's motivation to contribute to the cause. If more people think and write about worst-case futures, we will make greater progress on the crucial question of how to best avert s-risks.

The next step could be to think carefully about what you do in your working life. After all, few decisions are as consequential as our choice of career.[230] If you are interested in dedicating your career to the cause, you could consider applying for a role at an organisation that contributes to s-risk reduction.

Two research organisations with an explicit s-risk focus are the Center for Reducing Suffering (which I co-founded) and the Center on Long-Term Risk. Many other groups also do valuable work that contributes indirectly to s-risk reduction, even if it is not their primary or explicit goal. For example, the Organisation for Prevention of

---

[227] For more details on the advantages and drawbacks of research, see Vinding, 2022b.

[228] In fact, one could argue that the abstract nature of research or capacity building implies that these activities are likely to be particularly neglected.

[229] For instance, you could join the Facebook group "Risks of Astronomical Suffering (s-risks)".

[230] See also Baumann, 2022c.

Intense Suffering (OPIS) works to increase concern for suffering, and the work of Animal Ethics and Sentience Institute helps to expand humanity's moral circle to include all sentient beings.

Another option is to donate to such organisations and thereby support their important work.

Finally, perhaps the most important thing to do is to learn more about s-risks and related subjects. It is crucial that we invest in developing our future skills and abilities. After all, our skills and our level of knowledge largely determine how effectively we can improve the world.[231]

[231] For more details on these points, see Vinding, 2020a, Chapters 15-18.

Acknowledgments

The support of my colleagues at the Center for Reducing Suffering (Magnus Vinding, Teo Ajantaival, and Winston Oswald-Drummond) has been crucial in completing this book. Their feedback and comments have helped shape the book, for which I am very grateful.

Deep thanks go to Magnus Vinding in particular, for his encouragement and detailed input throughout the entire writing process. Magnus has probably influenced my views more than anyone else, and it is safe to say that this book would not have existed had it not been for his work. I can never thank him enough for his tireless work to reduce suffering.

I am also deeply grateful to Oscar Horta, Simon Knutsson, Jim Buhler, Jacy Reese, Jamie Harris, Caspar Oesterheld, David Althaus, Miranda Zhang, Annabella Wheatley, Timothy Chan, Imma Six, Michael St. Jules, Anthony DiGiovanni, Leonard Dung, and Dario Citrini for reading an early draft and providing useful comments. This feedback has helped to improve the book in many ways.

Bibliography

Achen, C.H., Bartels, L.M., 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government.* Princeton University Press.

Aird, M., 2020. Venn diagrams of existential, global, and suffering catastrophes. Retrieved from: *https://forum.effectivealtruism.org/posts/AJbZ2hHR4bmeZKznG/venn-diagrams-of-existential-global-and-suffering*

Ajantaival, T., 2021a. Positive roles of life and experience in suffering-focused ethics. Retrieved from: *https://centerforreducingsuffering.org/research/positive-roles-of-life-and-experience-in-suffering-focused-ethics/*

Ajantaival, T., 2021b. Minimalist axiologies and positive lives. Retrieved from: *https://centerforreducingsuffering.org/research/minimalist-axiologies-and-positive-lives/*

Ajantaival, T., 2022. Peacefulness, nonviolence, and experientialist minimalism. Retrieved from: *https://centerforreducingsuffering.org/research/peacefulness-nonviolence-and-experientialist-minimalism/*

Althaus, D. and Gloor, L., 2016. Reducing risks of astronomical suffering: a neglected priority. Retrieved from: *https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/*

Althaus, D., 2018. Descriptive Population Ethics and Its Relevance for Cause Prioritization. Retrieved from: *https://forum.effectivealtruism.org/posts/CmNBmSf6xtMyYhvcs/descriptive-population-ethics-and-its-relevance-for-cause*

Althaus, D., Baumann, T., 2020. Reducing long-term risks from malevolent actors. Retrieved from: *https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors*

Andrews, W., 2013. *Medieval Punishments: An Illustrated History of Torture.*

Animal Ethics, 2020. Wild animal suffering video course. Retrieved from: *https://www.animal-ethics.org/wild-animal-suffering-video-course/*

Anthis, J.R. and Paez, E., 2021. Moral circle expansion: A promising strategy to impact the far future. *Futures*, 130, p.102756.
Anthis, J.R., 2017. Summary of Evidence for Foundational Questions in Effective Animal Advocacy. Retrieved from: *https://www.sentienceinstitute.org/foundational-questions-summaries*
Bar-On, Y. et al., 2018. The biomass distribution on Earth. *PNAS*, 115(25), pp. 6506-6511
Bartels, D.M., 2006. Proportion dominance: The generality and variability of favoring relative savings over absolute savings. *Organizational Behavior and Human Decision Processes*, 100(1), pp.76-95.
Bartels, L.M., 1996. Uninformed votes: Information effects in presidential elections. *American journal of political science*, pp.194-230.
Bastardi, A., Uhlmann, E.L. and Ross, L., 2011. Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological science*, 22(6), p.731.
Bastian, B., Loughnan, S., Haslam, N. and Radke, H.R., 2012. Don't mind meat? The denial of mind to animals used for human consumption. *Personality and Social Psychology Bulletin*, 38(2), pp.247-256.
Baumann, T., 2017a. Arguments for and against moral advocacy. Retrieved from: *https://prioritizationresearch.com/arguments-for-and-against-moral-advocacy/*
Baumann, T., 2017b. Using surrogate goals to deflect threats. Retrieved from: *https://s-risks.org/using-surrogate-goals-to-deflect-threats/*
Baumann, T., 2018a. Research priorities for preventing threats. Retrieved from: *https://s-risks.org/research-priorities-for-preventing-threats/*
Baumann, T., 2018b. An introduction to worst-case AI safety. Retrieved from: *https://s-risks.org/an-introduction-to-worst-case-ai-safety/*
Baumann, T., 2018c. Focus areas of worst-case AI safety. Retrieved from: *https://s-risks.org/focus-areas-of-worst-case-ai-safety/*
Baumann, T., 2019a. Thoughts on longtermism. Retrieved from: *https://s-risks.org/thoughts-on-longtermism/*
Baumann, T., 2019b. How can we influence the long-term future? Retrieved from: *https://s-risks.org/how-can-we-influence-the-long-term-future/*
Baumann, T., 2020a. Common ground for longtermists. Retrieved from: *https://centerforreducingsuffering.org/research/common-ground-for-longtermi*

sts/
Baumann, T., 2020b. Thoughts on space colonisation. Retrieved from: *https://s-risks.org/thoughts-on-space-colonisation/*
Baumann, T., 2020c. Longtermism and animal advocacy. Retrieved from: *https://centerforreducingsuffering.org/longtermism-and-animal-advocacy/*
Baumann, T., 2020d. Representing future generations in the political process. Retrieved from: *https://centerforreducingsuffering.org/research/representing-future-generations-in-the-political-process/*
Baumann, T., 2020e. Space governance is important, tractable and neglected. Retrieved from: *https://forum.effectivealtruism.org/posts/QkRq6aRA84vv4xsu9/space-governance-is-important-tractable-and-neglected*
Baumann, T., 2021. Is most expected suffering due to worst-case outcomes? Retrieved from: *https://s-risks.org/wp-content/uploads/2021/02/Is_most_expected_suffering_due_to_worst_case_scenarios_.pdf*
Baumann, T., 2022a. How the animal movement could do even more good. Retrieved from: *https://centerforreducingsuffering.org/how-the-animal-movement-could-do-even-more-good/*
Baumann, T., 2022b. Five recommendations for better political discourse. Retrieved from: *https://centerforreducingsuffering.org/five-recommendations-for-better-political-discourse*
Baumann, T., 2022c. Career advice for reducing suffering. Retrieved from: *https://centerforreducingsuffering.org/research/career-advice-for-reducing-suffering/*
Baumann, T., Graepel, T. and Shawe-Taylor, J., 2020. Adaptive mechanism design: Learning to promote cooperation. In 2020 *International Joint Conference on Neural Networks* (IJCNN) (pp. 1-7).
Becerra, Ó., Johnson, N., Meier, P., Restrepo, J. and Spagat, M., 2012. Natural disasters, casualties and power laws: A comparative analysis with armed conflict. In *Proceedings of the annual meeting of the American Political Science Association.*
Beckstead, N., 2014. Will we eventually be able to colonize other stars?

Notes from a preliminary review. Retrieved from: *https://www.fhi.ox.ac.uk/will-we-eventually-be-able-to-colonize-other-stars-notes-from-a-preliminary-review/*
Beiser, Frederick C. 2016. *Weltschmerz: Pessimism in German Philosophy, 1860-1900.* Corby: Oxford University Press.
Bostrom, N., 2002. Existential risks. *Journal of Evolution and technology*, 9(1), pp.1-31.
Bostrom, N., 2003. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(3), pp.308-314.
Bostrom, N., 2013. *Anthropic bias: Observation selection effects in science and philosophy.* Routledge.
Bostrom, N., 2014. *Superintelligence: Paths, dangers, strategies.* Oxford University Press.
Bostrom, N., 2019. The vulnerable world hypothesis. *Global Policy*, 10(4), pp.455-476.
Braumoeller, B.F., 2019. *Only the dead: the persistence of war in the modern age.* Oxford University Press.
Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. and Anderson, H., 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.
Caplan, B., 2006. The totalitarian threat. Retrieved from: *https://essaydocs.org/the-totalitarian-threat.html*
Center for Reducing Suffering, 2021. Open Research Questions. Retrieved from: *https://centerforreducingsuffering.org/open-research-questions/*
Chalmers, D.J., 1995. Absent qualia, fading qualia, dancing qualia. In *Conscious experience* (pp. 309-328).
Christian, B., 2020. *The alignment problem: Machine learning and human values.*
Clifton, J., 2020. Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda. Retrieved from: *https://longtermrisk.org/research-agenda*
Cochrane, A., 2018. *Sentientist politics: A theory of global inter-species justice.* Oxford University Press.
Cohen, S., 2001. *States of denial: Knowing about atrocities and suffering.* John Wiley & Sons.
Cook, T., 2022. Replicating and extending the grabby aliens model.

Retrieved from: *https://forum.effectivealtruism.org/posts/7bc54mWtc7BrpZY9e/replicating-and-extending-the-grabby-aliens-model*

Cowen, T. and Parfit, D., 1992. Against the social discount rate. *Justice between age groups and generations*, 144, p.145.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K. and Graepel, T., 2021. Cooperative AI: machines must learn to find common ground. Retrieved from: *https://www.nature.com/articles/d41586-021-01170-0*

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K.R., Leibo, J.Z., Larson, K. and Graepel, T., 2020. Open problems in cooperative AI. arXiv preprint arXiv:2012.08630.

Deudney, D., 2020. *Dark Skies: Space Expansionism, Planetary Geopolitics, and the Ends of Humanity.* Oxford University Press.

Donovan, S.A., Labonte, M. and Dalaker, J., 2016. The US income distribution: Trends and issues. *Congressional Research Service Report.*

Dos Santos, T. R., 2020. *Why Not Parliamentarism?*

Drutman, L., 2020. *Breaking the Two-Party Doom Loop: The Case for Multiparty Democracy in America.* Oxford University Press.

Drutman, L., and Strano, M., 2021. What We Know About Ranked Choice Voting. Retrieved from: *https://www.newamerica.org/political-reform/reports/what-we-know-about-ranked-choice-voting/*

Duverger, M., 1954. *Political Parties: Their Organization and Activity in the Modern State.* Cambridge University Press.

Erdélyi, O.J. and Goldsmith, J., 2022. Regulating artificial intelligence: Proposal for a global solution. *Government Information Quarterly*, p.101748.

Fairvote Canada, 2018. A Look at the Evidence for Proportional Representation. Retrieved from: *https://www.fairvote.ca/2018/10/24/evidence/*

Fehige, Christoph. 1998. A Pareto Principle for Possible People. In *Preferences, 508–543.*

Fernbach, P.M., Rogers, T., Fox, C.R. and Sloman, S.A., 2013. Political extremism is supported by an illusion of understanding. *Psychological science*, 24(6), pp.939-946.

Freinacht, H., 2017. *The Listening Society: A Metamodern Guide to Politics, Book One.* Metamoderna.

Freinacht, H., 2019. *Nordic Ideology: A Metamodern Guide to Politics, Book Two.* Metamoderna.
Furnham, A., 2003. Belief in a just world: Research progress over the past decade. *Personality and individual differences*, 34(5), pp.795-817.
Garfinkel, B. and Dafoe, A., 2019. How does the offense-defense balance scale? *Journal of Strategic Studies*, 42(6), pp.736-763.
Gerrard, B., 2020. Are There Really More Slaves Now Than Anytime In History? Retrieved from: *https://braydeng.medium.com/are-there-more-slaves-now-than-anytime-in-history-38420e0542e5*
Gesicki, K., Zijlstra, A.A. and Miller Bertolami, M.M., 2018. The mysterious age invariance of the planetary nebula luminosity function bright cut-off. *Nature Astronomy*, 2(7), pp.580-584.
Gilens, M., 2001. Political ignorance and collective policy preferences. *American Political Science Review*, 95(2), pp.379-396.
Gloor, L., 2016a. Altruists Should Prioritize Artificial Intelligence. Retrieved from: *https://longtermrisk.org/altruists-should-prioritize-artificial-intelligence/*
Gloor, L., 2016b. The Case for Suffering-Focused Ethics. Retrieved from: *https://longtermrisk.org/the-case-for-suffering-focused-ethics/*
Gloor, L., 2016c. Suffering-focused AI safety: In favor of "fail-safe" measures. Retrieved from: *https://longtermrisk.org/files/fail-safe-ai.pdf*
Gloor, L., 2017. Tranquilism. Retrieved from: *https://longtermrisk.org/tranquilism/*
Gloor, L., 2018. Cause prioritization for downside-focused value systems. Retrieved from: *https://forum.effectivealtruism.org/posts/225Aq4P4jFPoWBrb5/cause-prioritization-for-downside-focused-value-systems*
Greaves, H. and MacAskill, W., 2019. The case for strong longtermism. *GPI Working Paper* No. 5-2021.
Hajjar, L., 2013. *Torture: A sociology of violence and human rights.* Routledge.
Handwerk, B., 2021. An Evolutionary Timeline of Homo Sapiens. Retrieved from: *https://www.smithsonianmag.com/science-nature/essential-timeline-understanding-evolution-homo-sapiens-180976807/*
Hannon, M., 2021. Disagreement or badmouthing? The role of expressive discourse in politics.

Hanson, R., 2011. A Galaxy On Earth. Retrieved from: *https://www.overcomingbias.com/2011/07/a-galaxy-on-earth.html*
Hanson, R., 2014a. Look Hard, Then Steer Slightly. Retrieved from: *https://fqxi.org/data/essay-contest-files/Hanson_FQXI_Essay.pdf*
Hanson, R., 2016. *The Age of Em: Work, Love, and Life when Robots Rule the Earth.* Oxford University Press.
Hanson, R., 2018. Vulnerable World Hypothesis. Retrieved from: *https://www.overcomingbias.com/2018/11/vulnerable-world-hypothesis.html*
Hanson, R., 2022. Why Not Wait On AI Risk? Retrieved from: *https://www.overcomingbias.com/2022/06/why-not-wait-on-ai-risk.html*
Hanson, R., Yudkowsky, E., 2013. The Hanson-Yudkowsky AI-Foom Debate. Machine Intelligence Research Institute, 2013. Retrieved from: *https://intelligence.org/files/AIFoomDebate.pdf*
Hanson, R., Martin, D., McCarter, C. and Paulson, J., 2021. A Simple Model of Grabby Aliens. arXiv e-prints, pp.arXiv-2102.
Harris, J., 2019. How Tractable is Changing the Course of History? Retrieved from: *https://www.sentienceinstitute.org/blog/how-tractable-is-changing-the-course-of-history*
Harris, J. and Anthis, J.R., 2021. The moral consideration of artificial entities: a literature review. *Science and engineering ethics*, 27(4), pp.1-95.
Henrich, J., 2015. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter.*
Horta, O., 2010. What is speciesism? *Journal of agricultural and environmental ethics*, 23(3), pp.243-266.
Horta, O., 2017. Animal Suffering in Nature: The Case for Intervention. *Environmental Ethics*, 39(3),pp.261-279.
Horta, O., 2022. *Making a Stand for Animals.* Routledge.
Howell, E., 2021. How many stars are in the Milky Way? Retrieved from:
https://www.space.com/25959-how-many-stars-are-in-the-milky-way.html
Howell, E., Harvey, A., 2022. How many galaxies are there? Retrieved from:
https://www.space.com/25303-how-many-galaxies-are-in-the-universe.html
Intergovernmental Panel on Climate Change, 2022. Climate Change 2022: Impacts, Adaptation and Vulnerability. Retrieved from: *https://www.ipcc.ch/report/ar6/wg2/*
Imai, K., and Lo, J., 2021. Robustness of Empirical Evidence for the

Democratic Peace: A Nonparametric Sensitivity Analysis. *International Organization*, 75(3), pp. 901-919.
Johannsen, K., 2017. Animal rights and the problem of r-strategists. *Ethical Theory and Moral Practice*, 20(2), pp.333-345.
Johnson, S. G. B., Merchant, T., & Keil, F. C., 2020. Belief digitization: Do we treat uncertainty as probabilities or as bits? *Journal of Experimental Psychology: General*, 149(8), 1417–1434.
Kahneman, D., 2011. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.
Knutsson, S., 2016. Measuring happiness and suffering. Retrieved from: *https://www.simonknutsson.com/measuring-happiness-and-suffering/*
Knutsson, S., 2022. Undisturbedness as the hedonic ceiling. Retrieved from:
https://www.simonknutsson.com/undisturbedness-as-the-hedonic-ceiling/
Kunda, Z., 1990. The case for motivated reasoning. *Psychological bulletin*, 108(3), p.480.
Lavie, C.J., Arena, R., Swift, D.L., Johannsen, N.M., Sui, X., Lee, D.C., Earnest, C.P., Church, T.S., O'Keefe, J.H., Milani, R.V. and Blair, S.N., 2015. Exercise and the cardiovascular system: clinical science and cardiovascular outcomes. *Circulation research*, 117(2), pp.207-219.
Lewis, G., 2016. Beware surprising and suspicious convergence. Retrieved from: *https://forum.effectivealtruism.org/posts/omoZDu8ScNbot6kXS/beware-surprising-and-suspicious-convergence*
Lodge, M. and Taber, C.S., 2005. The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology*, 26(3), pp.455-482.
Lukić, P. and Živanović, M., 2021. Shedding light on the Light Triad: Further evidence on structural, construct, and predictive validity of the Light Triad. *Personality and Individual Differences*, 178, p.110876.
MacAskill, W., 2015. *Doing good better: Effective altruism and a radical new way to make a difference.* Guardian Faber Publishing.
MacAskill, W., 2020. Are we Living at the Hinge of History? *Global Priorities Institute*.
MacAskill, W., 2022. *What We Owe the Future.* Basic Books.
Mata, A., 2016. Proportion dominance in valuing lives: The role of deliberative thinking. *Judgment & Decision Making*, 11(5), 441–449.
Mayerfeld, J., 1999. *Suffering and Moral Responsibility.* Oxford

University Press.
Mercier, H. & Sperber, D., 2017. *The Enigma of Reason.* Harvard University Press.
Metzinger, T., 2021. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(01), pp.43-66.
Minson, J.A. and Monin, B., 2012. Do-gooder derogation: Disparaging morally motivated minorities to defuse anticipated reproach. *Social Psychological and Personality Science*, 3(2), pp.200-207.
Moshagen, M., Hilbig, B. E., Zettler, I., 2018. The dark core of personality. *Psychological Review*, 125(5), 656–688.
Muehlhauser, L., 2017. Report on Consciousness and Moral Patienthood. Retrieved from: *https://www.openphilanthropy.org/research/2017-report-on-consciousness-and-moral-patienthood/*
Narveson, J., 1973. Moral problems of population. *The Monist*, pp.62-86.
Newman, M.E., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5), pp.323-351.
North, T.C., McCullagh, P., Tran, Z.V., Lavallee, D.E., Williams, J.M., Jones, M.V. and Papathomas, A.C., 2008. Effect of exercise on depression.
Nozick, R., 1974. *Anarchy, State, and Utopia.* Basic Books.
Oesterheld, C., 2017. Multiverse-wide cooperation via correlated decision making. *Foundational Research Institute.*
Oesterheld, C. and Conitzer, V., 2022. Safe Pareto improvements for delegated game playing. *Autonomous Agents and Multi-Agent Systems*, 36(2), pp.1-47.
Ord, T., 2015. Moral Trade. *Ethics*, 126, pp. 118-138. Retrieved from: *https://www.fhi.ox.ac.uk/wp-content/uploads/moral-trade-1.pdf*
Paulhus, D. L., 2014. Toward a taxonomy of dark personalities. *Current Directions in Psychological Science*, 23(6), 421-426.
Pearce, D., 1995. *The Hedonistic Imperative.*
Pinker, S., 2011. *The better angels of our nature: Why violence has declined.* Viking Books.
Planck Collaboration, 2020. Planck 2018 results-VI. Cosmological parameters. *Astronomy & Astrophysics*, 641, p.A6.
Plous, S., 1993. *The Psychology of Judgment and Decision Making*. New York: McGraw-Hill.

Ritchie, H., 2019. The world now produces more seafood from fish farms than wild catch. Retrieved from: *https://ourworldindata.org/rise-of-aquaculture*
Ritchie, H., Roser, M., 2017. Meat and Dairy Production. Retrieved from:
https://ourworldindata.org/meat-production#number-of-animals-slaughtered
Roser, M., Ortiz-Ospina, E. and Ritchie, H., 2013. Life expectancy. *Our World in Data.*
Sandberg, A., Drexler, E. and Ord, T., 2018. Dissolving the Fermi paradox. arXiv preprint arXiv:1806.02404.
Schonfeld, Bryan, 2020. Democracy Promotion as an EA Cause Area. Retrieved from: *https://forum.effectivealtruism.org/posts/dTconqCtsmHQsNwo9/democracy-promotion-as-an-ea-cause-area-1*
Schukraft, J., 2019. Invertebrate Welfare Cause Profile. Retrieved from: *https://forum.effectivealtruism.org/posts/EDCwbDEhwRGZjqY6S/invertebrate-welfare-cause-profile*
Schwartz, S., 2022. Ben Franklin's gift that keeps on giving. Retrieved from: *https://www.historynet.com/ben-franklins-gift-keeps-giving/*
Scully, M., 2002. *Dominion: The power of man, the suffering of animals, and the call to mercy.* Macmillan.
Sentience Institute, 2017. Animals, Food, and Technology (AFT) Survey 2017. Retrieved from: *https://www.sentienceinstitute.org/animal-farming-attitudes-survey-2017*
Shulman, C., Bostrom, N., 2014. Embryo selection for cognitive enhancement: curiosity or game-changer? *Global Policy*, vol. 5, pp. 85–92.
Simler, K., Hanson, R., 2017. *The Elephant in the Brian: Hidden Motives in Everyday Life.* Oxford University Press.
Singer, P., 1975. *Animal liberation: A New Ethics for Our Treatment of Animals.*
Slovic, P., 2010. If I look at the mass I will never act: Psychic numbing and genocide. In *Emotions and risky technologies* (pp. 37-59). Springer, Dordrecht.
Sotala, K. and Gloor, L., 2017. Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica*, 41(4).
Sotala, K., 2012. Advantages of artificial intelligences, uploads, and digital minds. *International journal of machine consciousness*, 4(01),

pp.275-291.
Sotala, K., 2017. How feasible is the rapid development of artificial superintelligence? *Physica Scripta*, *92*(11), p.113001.
Tarsney, C., 2019. The epistemic challenge to longtermism. *Global Priorities Institute.*
Taylor, S., 2019. Pathocracy. Retrieved from: *https://www.psychologytoday.com/us/blog/out-the-darkness/201907/pathocracy*
Thompson, S., 1999. Illusions of Control: How We Overestimate Our Personal Influence. *Current Directions in Psychological Science*, 8(6), pp. 187-190.
Tomasik, B., 2006. On the Seriousness of Suffering. Retrieved from: *https://reducing-suffering.org/on-the-seriousness-of-suffering/*
Tomasik, B., 2007. Why Maximize Expected Value? Retrieved from: *https://reducing-suffering.org/why-maximize-expected-value*
Tomasik, B., 2009. How Many Wild Animals Are There? Retrieved from: *https://reducingsuffering.org/how-many-wild-animals-are-there/*
Tomasik, B., 2011. Risks of astronomical future suffering. Retrieved from: *https://longtermrisk.org/files/risks-of-astronomical-future-suffering.pdf*
Tomasik, B., 2013a. Gains from Trade through Compromise. Retrieved from: *https://longtermrisk.org/gains-from-trade-through-compromise/*
Tomasik, B., 2013b. Does the Animal-Rights Movement Encourage Wilderness Preservation? Retrieved from: *https://reducing-suffering.org/does-the-animal-rights-movement-encourage-wilderness-preservation/*
Tomasik, B., 2013c. Omelas and Space Colonization. Retrieved from: *https://reducing-suffering.org/omelas-and-space-colonization/*
Tomasik, B., 2014. Reasons to Be Nice to Other Value Systems. Retrieved from: *https://longtermrisk.org/reasons-to-be-nice-to-other-value-systems/*
Tomasik, B., 2015a. Should Altruists Focus on Reducing Short-Term or Far-Future Suffering? Retrieved from: *https://reducing-suffering.org/altruists-focus-reducing-short-term-far-future-suffering/*
Tomasik, B., 2015b. Against Wishful Thinking. Retrieved from: *https://longtermrisk.org/against-wishful-thinking/*
Tomasik, B., 2015c. Reasons to Promote Suffering-Focused Ethics. Retrieved from:

https://reducing-suffering.org/the-case-for-promoting-suffering-focused-ethics/

Tomasik, B., 2017. Will Future Civilization Eventually Achieve Goal Preservation? Retrieved from: *https://reducing-suffering.org/will-future-civilization-eventually-achieve-goal-preservation/*

Torres, P., 2018a. Space colonization and suffering risks: Reassessing the "maxipok rule". *Futures*, 100, pp. 74-85.

Torres, P., 2018b. Why We Should Think Twice About Colonizing Space. *Nautilus*. Retrieved from: *https://nautil.us/why-we-should-think-twice-about-colonizing-space-237149/*

Trammell, P., 2021. Dynamic Public Good Provision under Time Preference Heterogeneity: Theory and Applications to Philanthropy. *Global Priorities Institute.*

Tversky, A. and Kahneman, D., 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2), pp.207-232.

Tversky, A. and Kahneman, D., 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), pp.1124-1131.

Vinding, M., 2015. *Speciesism: Why It Is Wrong and the Implications of Rejecting It.*

Vinding, M., 2016a. *Reflections on Intelligence.*

Vinding, M., 2016b. Animal Advocates Should Focus On Antispeciesism, Not Veganism. Retrieved from: *https://magnusvinding.com/2020/05/04/animal-advocates-should-focus-on-antispeciesism-not-veganism/*

Vinding, M., 2017. The future of growth: near-zero growth rates. Retrieved from: *https://magnusvinding.com/2020/05/04/the-future-of-growth-near-zero-growth-rates/*

Vinding, M., 2018a. Moral Circle Expansion Might Increase Future Suffering. Retrieved from: *https://magnusvinding.com/2018/09/04/moral-circle-expansion-might-increase-future-suffering/*

Vinding, M., 2018b. In Defense of Nuance. Retrieved from: *https://magnusvinding.com/2018/08/29/in-defense-of-nuance/*

Vinding, M., 2018c. Why Altruists Should Perhaps Not Prioritize

Artificial Intelligence: A Lengthy Critique. Retrieved from: *https://magnusvinding.com/2018/09/18/why-altruists-should-perhaps-not-prioritize-artificial-intelligence-a-lengthy-critique/*
Vinding, M., 2018d. *Effective Altruism: How Can We Best Help Others?* Ratio Ethica.
Vinding, M., 2020a. *Suffering-Focused Ethics: Defense and Implications.* Ratio Ethica.
Vinding, M., 2020b. Why altruists should be cooperative. Retrieved from: *https://centerforreducingsuffering.org/research/why-altruists-should-be-cooperative/*
Vinding, M., 2020c. On fat-tailed distributions and s-risks. Retrieved from: *https://centerforreducingsuffering.org/on-fat-tailed-distributions-and-s-risks/*
Vinding, M., 2020d. Ten Biases Against Prioritizing Wild-Animal Suffering. Retrieved from: *https://magnusvinding.com/2020/07/02/ten-biases-against-prioritizing-wild-animal-suffering/*
Vinding, M., 2022a. *Reasoned Politics.* Ratio Ethica.
Vinding, M., 2022b. Research vs. non-research work to improve the world: In defense of more research and reflection. Retrieved from: *https://magnusvinding.com/2022/05/09/in-defense-of-research/#the-case-for-more-research*
Vinding, M., 2022c. A phenomenological argument against a positive counterpart to suffering. Retrieved from: *https://centerforreducingsuffering.org/phenomenological-argument/*
Vinding, M., 2022d. What does a future dominated by AI imply? Retrieved from: *https://magnusvinding.com/2022/09/06/what-does-a-future-dominated-by-ai-imply/*
Vinding, M., 2022e. Radical uncertainty about outcomes need not imply (similarly) radical uncertainty about strategies. Retrieved from: *https://magnusvinding.com/2022/09/07/strategic-uncertainty/*
Vinding, M., 2022f. Beware underestimating the probability of very bad outcomes: Historical examples against future optimism. Retrieved from: *https://magnusvinding.com/2022/09/09/beware/*
Vinding, M., Baumann, T., 2021. S-risk impact distribution is double-tailed. Retrieved from:

https://centerforreducingsuffering.org/s-risk-impact-distribution-is-double-tailed/

West, B., 2017. An Argument for Why the Future May Be Good. Retrieved from: *https://forum.effectivealtruism.org/posts/kNKpyf4WWdKehgvRt/an-argument-for-why-the-future-may-be-good*

White, M., 2011. *Atrocities: The 100 deadliest episodes in human history.* WW Norton & Company.

Williams, E.G., 2015. The possibility of an ongoing moral catastrophe. *Ethical theory and moral practice*, 18(5), pp.971-982.

Yudkowsky, E., 2013. Pascal's Muggle: Infinitesimal Priors and Strong Evidence. Retrieved from: *https://www.lesswrong.com/posts/Ap4KfkHyxjYPDiqh2/pascal-s-muggle-infinitesimal-priors-and-strong-evidence*