Uncertainty about the expected moral value of the long-term future: is reducing human extinction risk valuable?

MA Thesis Siebe Rozendal (2192667)

Philosophy of a specific discipline (Strategic Innovation Management)

31 - 10 - 2019

Word count: 19,451
(references & footnotes excluded)

Supervisor:	dr. Ryan Doody, University of Groningen
Second supervisor:	dr. Simon Beard, University of Cambridge
Additional assessor:	dr. Barteld Kooi, University of Groningen

university of groningen

faculty of philosophy

Abstract

The future can be enormously valuable if it contains a large population of sentient beings. Multiple authors have noted that, to improve the value of the long-term future, we should reduce the risk of human extinction. This would increase the expected size of the future population, which is good if one expects the average moral value of a life to be positive. However, a future without human extinction is not necessarily more valuable than a future in which humanity goes extinct. To assess the value of reducing human extinction risk, I first look at the expected moral value of the long-term future. Moral uncertainty between Totalism and Asymmetric Views makes it difficult to assess the expected moral value of a single possible future. However, even if the future would be worse than extinction, reducing the risk of human extinction risk could still be positive; reducing the sources of human extinction risk will also reduce the risk of global catastrophe. I argue that we should expect global catastrophes to have a negative effect on the value of the long-term future if they do not lead to extinction. If global catastrophes are unlikely to lead to extinction, this would be a reason in favour of reducing the sources of extinction risk. In conclusion, the expected moral value of human extinction risk reduction depends on one's moral uncertainty between Totalism and Asymmetric Views and on the likelihood of global catastrophes to lead to extinction.

Key Words*: longtermism; moral uncertainty; extinction risk; existential risk; global catastrophic risk; expected moral value*

Disclaimer

This thesis discusses topics of astronomical scale, the value of humanity's continued existence, and the badness of (extreme) suffering. Even discussed in relatively dry analytical style, it is still possible that these topics cause strong emotions when they are contemplated on deeply. I would encourage readers to keep this in mind, to realize that these emotions are okay to experience when they arise, and to treat them with compassion.

I will always reply to questions, concerns, and ideas about this thesis, although I cannot guarantee that my reply will be helpful. I especially encourage anyone who wants to take action based on this thesis to contact me and I will try to help.

You can contact me on sieberozendal@gmail.com or through sieberozendal.com.

Acknowledgements

I would like to thank Ryan Doody for his generous offer of time and helpful comments in supervising this thesis. Without him, this thesis would not have existed. I would also like to thank Simon Beard for offering initial advice on a thesis topic and providing feedback throughout the process. In addition, I would like to thank Justin Timmer, Rick Kloosterman, and Sara Elsholz who, as good friends, dutifully proofread this thesis outside their area of expertise. Lastly, I am grateful for all the stimulating discussions I have had about the topics in this thesis with people from the Effective Altruism community.

Contents

Abstract 1

Chapter 1. Introduction 4

1.1 The longtermism paradigm 4

1.2 Focus on reducing extinction risk 6

1.3 The expected moral value of the long-term future 8

1.4 How should the EMV of the long-term future be assessed? 8

1.5 Focus on ‘astronomical’ futures 15

Chapter 2. Problems with assessing the EMV of a possible -future 16

2.1 A theoretical method to deal with moral uncertainty: expected moral value 16

2.1.1 Introducing expected moral value 16

2.1.2 The problem of intertheoretic value comparison 18

2.2 Does Totalism dominate the EMV for large population cases? 19

2.2.1 In large population cases, unbounded moral theories overwhelm others 19

2.2.2 Totalism is not the only unbounded view; enter Asymmetric Views 20

2.3 Can the EMV approach resolve (big) disagreement? 25

2.3.1 Disagreement and the problem of intertheoretic value comparison 25

2.3.2 Big worlds breed big disagreement 32

Chapter 3. Problems with assessing the EMV of the entire long-term future 35

3.1 Assessing the EMV of the entire long-term future introduces empirical uncertainty35

3.2 Strong empirical uncertainty makes the EMV fragile 36

Chapter 4. Given our uncertainties, is it good to reduce human extinction risk? 40

4.1 Pragmatic methods to deal with moral uncertainty 40

4.1.1 Option value: neither clearly positive, nor big 41

4.1.2 We cannot wait to resolve moral uncertainty and disagreement 44

4.1.3 Doing nothing is also a choice 44

4.2 Does reducing extinction risk have positive expected value? 45

4.2.1 Focus on the EMV of trying to reduce human extinction risk 45

4.2.2 A primer on extinction risks 45

4.2.3 Positive side-effects of reducing non-AI sources of extinction risk 47

Chapter 5. Conclusion 52

5.1 Summary & conclusion 52

5.2 Theoretical implications 53

5.3 Practical implications 53

5.4 Limitations and further research 55

References 57

Chapter 1. Introduction

Reading ethics, one may get the impression that there are few topics on which moral philosophers agree. However, most moral theories would agree that, if the future is significantly positive, it is better for humanity to continue to exist than to go extinct.[1] Rights-based theorists appeal to a need for intergenerational justice (Meyer, 2017), contractarians argue that a concern for future generations is the solution to an intergenerational collective action problem (Gardiner, 2009), and virtue ethicists base their concern for future generations on "the principle of benevolence and a recognition of the dignity and worth of all life" (Gaba, 1999, p. 252).[2] In addition, most people would answer they would like to prevent humanity's extinction if humanity had a long and glorious future (Caviola & Schubert, Unpublished raw data). Perhaps most naturally, nearly all consequentialist theories care about the long-term future of humanity. This has at least two reasons; first, most consequentialist theories hold that *all* consequences of acts are morally relevant, regardless of how, when, or where they occur. Second, most consequentialist theories lean towards impartialism regarding one's altruistic preferences: in benefiting others, we should not discriminate based on irrelevant moral characteristics, such as gender, skin color, species, location, and - some theories would include - time and modality of existence. This thesis looks at the philosophical issues around human extinction from the perspective of consequentialist theories.

1.1 The longtermism paradigm

Recently, a number of (mostly consequentialist) philosophers have argued that not only is the long-term future an important moral concern, focusing on the long-term future should be humanity's *primary* concern (Beckstead, 2013; Bostrom, 2003; Parfit, 1984). As phrased by Beckstead (2013, p. 1):

> From a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

[1] However, there is always an exception. Some theories imply *anti-natalism*: that it cannot be good for someone, or the whole of humanity, to exist, but it can be bad (Benatar, 2008). These views will be discussed in section 3.4.

[2] I should note that, in my perspective, many of these theories needed to be expanded to incorporate the intuition that future generations matter, rather than that it was a natural conclusion of these theories. For example, the original set-up of the social contract involves a single-generation contract (Gardiner, 2009).

We call this the *longtermism paradigm.* It is driven by at least the following two reasons. First, the possible future of Earth-originating life is incredibly vast. Timewise, the universe may be habitable up to 100,000,000,000,000 years (Beckstead, 2013).[3] Even with the current population size of Earth, and assuming a lifespan of 100 years, that would amount to 7,000,000,000,000,000,000,000 human lives being lived. However, given that the reachable universe is much larger than Earth, there are resources available to support far larger population sizes which I will not estimate here. Because those futures with large population sizes are not improbable, the *expected* length of the future and the *expected* amount of lives being lived are large.

Second, we are currently able to influence the long-term future significantly: some actions we can currently take may drastically influence humanity's future trajectory. For example, our influence on Earth's climate affects many future generations, and the specific designs of breakthrough technologies and legal documents (e.g. a constitution of a possible world government) can have persisting effects in the long-term. Another lever to affect humanity's future trajectory is our influence on the risk of human extinction. As humanity's technological power increases, we not only increase our capability to improve the world, but also increase our capability to destroy it (Bostrom, 2002; Leslie, 2002). An informal survey among 13 experts put the risk of extinction before 2100 at 19% (Sandberg & Bostrom, 2008), with the largest amount of risk stemming from emerging technologies. While these estimates should be taken with a substantial grain of salt - the probability of human extinction is hard to extrapolate from past evidence - there is an emerging consensus that the level of risk is not negligible (cf. Rowe & Beard (2018) for an examination of the evidence base for these assessments, as well as an overview of estimates). Our actions today can affect whether we develop powerful emerging technologies safely or callously. Given the stakes, 'safety first' seems the wisest course of action.

Although the two reasons above are sufficient to support longtermism, a third consideration increases its importance. It is the premise that it is good to bring new life into the universe when the life is sufficiently worth living. After all, would an empty universe not be so much worse than a universe filled with happy, thriving lives, with beauty, and with meaning?[4] Not to mention that these lives could have a vastly higher quality than

[3] This is based on how long stars will continue to burn. So-called 'white dwarfs' are estimated to last as long as 100 trillion years (Adams, 2008). However, it is not clear whether this is the limiting factor, or whether something more futuristic is the limiting factor. One possible hard limit is the amount of computations that can be made, which allows for even much longer and more subjective time, *if* consciousness can be computed.

[4] In most of this thesis I will refer to lives or well-being as what constitutes value in a world, but this does not mean I exclude other forms of intrinsic value. I believe the core of the thesis applies to many intrinsic values, as it seems many of them scale with population size: a larger population means there is more beauty to be created,

current conditions permit; with enough time and effort we could develop advanced technology to eliminate diseases, improve conditions for social relations to thrive, and more easily induce various states of bliss and enlightenment (Pearce, 1995).

1.2 Focus on reducing extinction risk

Assuming that what matters most is the long-term future, there are two main ways of altruistic impact. First, one can increase the *expected quality of life per person* in the future. For example, one can try to increase the probability that the future will contain cooperative and altruistic agents with a concern for all that is valuable, which will make it more likely that lives in the future will be worth living.

A second way of altruistic impact is to increase the *expected size* of the future: the expected amount of value-bearers in the future. For example, if one values aesthetic quality, increasing the expected size means increasing the expected amount of works of art and other objects that can have aesthetic quality. The most obvious way to increase the expected size of the future is to increase the probability that humanity will survive, expand, and progress. In this vain, different philosophers have come to similar but slightly different decision rules to follow. Bostrom (2013) uses *Maxipok:*

> Maximise the probability of an 'OK outcome', where an OK outcome is any outcome that avoids existential catastrophe.

Existential catastrophe needs to be defined further. Bostrom (2013, p. 15) mentions it in his definition of *existential risk*:

> An existential risk is one that threatens to cause the extinction of Earth-originating intelligent life or the permanent and drastic failure of that life to realise its potential for desirable development.

There are two problems with this definition. First, it includes only other scenarios that *permanently* curtail the desirable development of Earth-originating intelligent life. Besides extinction, Bostrom (2009) surveys other options like flawed realization, plateauing, and recurrent collapse. However, some events should still be classified as an existential catastrophe even if they do not curtail the future potential *for certain*; a drastic drop in the

more meaning to be experienced, and more knowledge to have. Furthermore, well-being seems the least controversial intrinsic value to have, such that most theories would care at least about well-being.

probability of realizing this potential is still catastrophic. Therefore, Cotton-Barratt and Ord (2015, p. 2) propose a different definition:

> An existential catastrophe is an event which causes the loss of a large fraction of expected value.

An example is a technologically advanced totalitarian regime that takes power; this drastically curtails future potential if the regime is stable and effective in preventing organized resistance (Caplan, 2008).[5], [6] Even if there is a tiny chance that Earth-originating intelligent life turns onto a positive trajectory eventually, this should probably still be classified as existential risk. On this second definition preventing existential catastrophe is, by its very definition, positive in expected value. 'Should we prevent existential catastrophe?' becomes a non-question.

Nonetheless, 'should we prevent human extinction?' remains an important question. In the same report Cotton-Barratt and Ord state "we certainly want to include all extinction events" (Cotton-Barratt & Ord, 2015, p. 1). However, extinction[7] events are not necessarily negative in expected value; it depends on whether the alternative of non-extinction is more valuable. A trajectory devoid of value is still better than a trajectory containing mostly suffering. As a result, the Cotton-Barratt & Ord definition, perhaps unintendedly, wisely remains silent on whether extinction events are existential catastrophes when there is no further information about the expected value of the future.

1.3 The expected moral value of the long-term future

What should we think about the prospect of extinction? The pessimist will say that we should welcome extinction because the future looks bleak. The optimist will say that we should wish for preservation because the future looks bright. However, both agree that the value of extinction depends on the value of the counterfactual: what would happen otherwise?

As a first step, we will compare the value of extinction to the value of the long-term future. More specifically, we will compare it to the *expected moral value (EMV)* of the long-term future. I will explain EMV in more detail in chapter 2. For now, the EMV of the

[5] This might become more likely with advanced surveillance technology, or when new minds (with a biological or artificial substrate) can be engineered by its creator: rebellious individuals may not be created.

[6] Also note that if such a regime has ambitions to expand into space, things could be far worse than 'curtailing humanity's potential'.

[7] From here onwards, whenever I speak of extinction I mean *human* extinction, because humans are the most likely group of actors to shape the long-term future.

long-term future is, simply put, the value we expect the future to have given that we are morally and empirically uncertain. Here, 'morally uncertain' means that we are uncertain about how to assign a moral value to the future even if we knew all the morally-relevant facts. Empirical uncertainty means that we are uncertain about the morally-relevant facts; we do not know what the future looks like.

To be accurate, we should not compare the value of extinction to the EMV of the long-term future. Instead, the value of extinction should actually be compared to the EMV of the long-term future *conditional on non-extinction*.[8] Solely for simplicity's sake we assume these to be similar enough to not affect the practical implications. In any case, the main part of this thesis discusses theoretical problems that apply to either way of specifying the EMV of the long-term future. Furthermore, it is a fascinating topic on its own. It continues in the tradition of grand narratives about where humanity is going: downward or upward?

1.4 How should the EMV of the long-term future be assessed?

It is an ambitious project to assess the expected moral value of the long-term future. To formalize the problem, I have to reduce the scope of the problem by making some assumptions. I justify some of the assumptions. Some other assumptions are merely for simplicity's sake.

Step 1. Partitioning the future with welfare-based sets

The first goal is to carve up the future into a partition: a set of elements that are *jointly exhaustive* (i.e. the sum of all elements covers the entire possibility space) and elements that are *mutually exclusive* (i.e. none of the elements overlap; if one element is 'true', all other elements must be 'false'). We need this partition in order to assign *credences* (i.e. subjective probabilities) to each possible future in a valid way; we want the sum of all credences to equal 1.

To create a morally-relevant partition of all possible futures, I define the following four sets. This partitioning is specific to the problems I want to address in this thesis. It cannot be used for all cases of moral uncertainty about the expected moral value of the future. This partitioning is insensitive to a number of moral questions.[9] In this thesis, I want

[8] The value of the future, conditional on non-extinction, might be different if the non-extinction is evidence that should affect our credences about what the future will be like. The condition of non-extinction might indicate that humanity would have overcome egoism, tribalism, war, and near-sightedness. Thus, the future might look substantially better if we assume humanity does not go extinct.

[9] This partitioning is optimized for the problems I want to address in this thesis. It cannot be used for all cases of moral uncertainty about the expected intertheoretic value of the future. This partitioning is insensitive to a

to address moral uncertainty about how to weigh lives with net-negative welfare to lives with net-positive welfare. Therefore, I partition the future according to how many lives are lived with net-positive welfare, how many lives are lived with net-negative welfare, the average welfare of the lives lived with net-positive welfare, and the average welfare of the lives lived with net-negative welfare. To do this, I will define a life as ‘a person’s sequence of experiences from birth until death’ (for debate on personal identity see Olson, 2019).

First, let Q^+ represent a finite set of possible total quantities of future lives with a net-positive average lifetime welfare. These are all the possibly correct answers to the question “How many lives will be lived from now on with a welfare more positive than no experience?” To avoid issues with infinite ethics, I assume that it is impossible to have an infinite amount of lives in the future.

$$Q^+ = \{q_1{}^+, q_2{}^+, q_3{}^+, \dots , q_n{}^+ : q_k{}^+ \geq 0\}$$

Second, let Q^- represent a finite set of possible total amounts of future lives with a net-negative average lifetime welfare. These are all the possibly correct answers to the question “How many lives will be lived from now on with a welfare worse than no experience?”

$$Q^- = \{q_1{}^-, q_2{}^-, q_3{}^-, \dots , q_n{}^- : q_k{}^- \geq 0\}$$

Third, let $\overline{W^+}$ represent a finite set of possible average lifetime welfare of a total future population with net-positive average lifetime welfare. This set is finite, and I assume welfare to be discrete, as well as finite.[10] By dividing up the possible future populations on the basis of their welfare, I assume that there is a non-arbitrary cutoff point where welfare is neutral and assign that point a value of zero welfare. I do not need to define where this point lies, although I stipulate that neutral/zero welfare is equivalent to non-existence. Therefore, again only for simplicity’s sake, I assume throughout this thesis that non-existence is comparable to existence, in contrast to some philosophers.[11]

$$\overline{W^+} = \{\overline{w_1{}^+}, \overline{w_2{}^+}, \overline{w_3{}^+}, \dots , \overline{w_n{}^+} \ : \ 0 \ \leq \ \overline{w_k{}^+} \ < \ \infty\}$$

number of things; 1) how welfare is distributed over time 2) how welfare is distributed over people, and 3) other properties than welfare that might be morally-relevant.

[10] For welfare to be finite, I assume that 1) the ‘amplitude’ of welfare cannot be infinite at any particular moment in time, and 2) a life can only have a finite duration of welfare.

[11] To say that non-existence is comparable to existence means that statements like “she would have been better off is she never had existed” make sense. See McMahan (2009) for a defense of non-comparability.

Fourth, let $\overline{W^-}$ represent a finite set of possible average lifetime welfare of a total future population with net-negative average lifetime welfare.

$$\overline{W^-} = \{\overline{w_1^-}, \overline{w_2^-}, \overline{w_3^-}, \dots, \overline{w_n^-} \ : \ -\infty < \overline{w_k^-} \leq 0\}$$

With these four sets, we can carve up the space of all possible futures into possible *axiological* futures (henceforth "possible α-futures"), such that each possible future is part of one and only one possible α-future. This simply means that we only look at the axiological properties of a future. When we describe the axiological properties of a future, this can refer to multiple possible futures that are different in their non-axiological properties. Let us call the set of all possible α-futures F. It is the Cartesian product of the sets above, i.e.

$$F = Q^+ \times \overline{W^+} \times Q^- \times \overline{W^-} = \{F_1, F_2, F_3, \dots, F_m\}$$

Each element F_j of this Cartesian product is an ordered quadruple. Because all sets of this product are finite, the Cartesian product F is also finite. We represent the number of elements of F by m. It is the product of the number of elements in all the subfactors of *F;*

$$m = n(Q^+) * n(\overline{W^+}) * n(\overline{W^-}) * n(Q^-) .$$

To illustrate these sets, let's apply them to an example. Imagine a possible α-future F_{110} with the following lives:

Life	Average welfare/year	Years lived	Total welfare of life
A	+10	80	+800
B	+0.5	100	+50
C	-5	70	-350
D	+11	100	+1,100
E	- 20	50	-1,000
F	-2.5	80	-200
G	+7	120	+840
H	+2	5	+10
I	- 15	90	- 1,350
J	+25	200	+ 5,000

Table 1. Lives in possible α-future F_{110} with their welfare levels.

This world F_{110} contains six lives with net-positive welfare (A, B, D, G, H, and J), so $q^{+}(F_{110}) = 6$. It contains four lives with net-negative welfare (C, E, F, and I), so $q^{-}(F_{110}) = 4$. The average welfare of the set of lives with net-positive welfare, denoted by $\overline{w^{+}(F_{110})} = \frac{800+50+1,100+840+10+5,000}{6} = 1,300$. The average welfare of the set of lives with net-negative welfare, denoted by $\overline{w^{-}(F_{110})} = \frac{-350-1,000-200-1,350}{4} = -725$. Summarising, $F_{110} = \left(q^{+}(F_{110}),\ \overline{w^{+}(F_{110})}, q^{-}(F_{110}),\ \overline{w^{-}(F_{110})}\right) = (6,\ 1300,\ 4,\ -725)$. The total welfare in this world, $W(F_{110})$ equals $6 * 1,300 + 4 * \ -725\ =\ 4,900$.

Step 2. Credence distribution over the partition

If we want to know the expected welfare of the future, we need to incorporate our empirical uncertainty about which α-future will actualise (i.e. which element of F will turn out to describe the entire future most accurately). Let $Cr(F)$ be a credence mass function over the space of possible α-futures. We distribute our credences $Cr(F_j)$ over the power set of F, $P(F)$ (i.e. all subsets of F), such that each subset (including each element) of F is assigned a credence in the interval $[0, 1]$.[12]

$$Cr(F):\ P(F) \to [0, 1]$$

[12] Note that, theoretically, our credences can affect how the future develops. which could make the credence not well-defined. For example, a prediction that the future looks bright could make people complacent, which could hurt our prospects. Therefore, for simplicity's sake, we assume that the credence function cannot influence real world affairs.

Let X represent a subset of F. Then the following conditions apply:

(1) For any $X \subseteq F$, $0 \leq Cr(X) \leq 1$,

(2) $Cr(F) = 1$,

(3) If X and Y are disjoint, $Cr(X \vee Y) = Cr(X) + Cr(Y)$

With this credence distribution, we can define a function for the *expected welfare* in the future:

$$EW(F) = \sum_{j=1}^{m} Cr(F_j) * W(F_j), \qquad \text{where}$$

$$W(F_j) = q^{+}(F_j) * \overline{w^{+}(F_j)} + q^{-}(F_j) * \overline{w^{-}(F_j)}$$

For a possible α-future F_j, this formula multiplies the amount of people that will have positive welfare in that world $q^{+}(F_j)$ times the average welfare of those people $\overline{w^{+}(F_j)}$, and adds the product of people with negative welfare times their average welfare. Note that this second product $q^{-}(F_j) * \overline{w^{-}(F_j)}$ is never positive, because $\overline{w^{-}(F_j)}$ is non-positive by stipulation.

However, we cannot directly translate the expected welfare to the expected moral value of the future. Different theories make different translations from welfare to moral value.

Step 3. Assigning moral value to possible α-futures

To obtain the moral value of a possible α-future, we need to define a value function V. This function will map all the possible α-futures onto the real line. Formally, $V : F \mapsto R$.[13]

$$V_{T_i}(F_j) = \textit{the value of a world } F_j, \textit{ according to theory } T_i$$

Different moral theories have different value functions. An example of a value function is Totalism[14], in which $V_{Tot.}(F_j) = 1 * q^{+}(F_j) * \overline{w^{+}(F_j)} + 1 * q^{-}(F_j) * \overline{w^{-}(F_j)}$.

[13] Since F is not continuous but a finite set, the value function V does not map to the complete range of the real line like it would if it were a linear transformation.

[14] Totalism is a well-respected axiology, in which the value of a world is simply the sum of all value and disvalue in that world, regardless of where it is located, when it is located, or how it is distributed. In addition, positive and negative welfare are weighed equally. This thesis assumes that welfare is the only property that matters.

Step 4. Assigning a credence function to possible value functions

We can now define the moral value of a possible α-future since we have a full description of the morally relevant properties in that possible α-future F_j and a value function V_{T_i}. However, we can be uncertain between which value function, which moral theory, is the correct one. To deal with this, we can define a credence function over moral theories, such that the set of moral theories is mutually exclusive and jointly exhaustive, and $0 \leq Cr(T_i) \leq 1$.

If we consider *n* moral theories, we can define the expected moral value of a single possible α-future with a formula based on the one used by MacAskill (2014, p. 17).

$$EMV(F_j) = \sum_{i=1}^{n} V_{T_i}(F_j) * Cr(T_i)$$

This formula calculates the weighted average moral value attributed to a possible α-future F_j. For a finite amount of moral theories T_i, a value is attributed to a possible α-future F_j. Next, these values are weighted by an agent's credence in the theory T_i. I will provide motivation for this EMV-approach in Chapter 2.

Step 5. Bringing everything together

We now have all the components required to define the expected moral value of the future. Based on the expected moral value of single possible α-futures, I define the expected moral value of the future by re-introducing the empirical uncertainty from step 2.

$$EMV(F) = \sum_{j=1}^{m} \sum_{i=1}^{n} V_{T_i}(F_j) * Cr(T_i) * Cr(F_j)$$

Each possible α-future F_j is assigned an expected moral value, and this value is then weighted by the credence that this possible α-future will actualize. All values are summed to arrive at the expected moral value of the future.

Therefore, the sum of all value in a world, according to Totalism, equals the sum of all welfare. The 1's in the formula indicate the equal weighting of positive and negative welfare, in contrast to Asymmetric views which will be discussed in Chapter 2.

1.5 Focus on ‘astronomical’ futures

There are large differences in the values that possible α-futures can take. Humanity, or something else of moral value, can persist for thousands, millions, billions, or even trillions of years. Not only that, humanity can either remain on one planet, or expand to the solar system, or expand to the galaxy, or to the local supercluster, or to all the stars within its reach. We can call these ‘astronomical’ α-futures, based on their potential to contain astronomical amounts of value (or disvalue!). Given these different scales, scenarios in which there is a lot of moral (dis)value will tend to dominate the EMV sum, unless the astronomical futures are significantly unlikely. The probability of astronomical futures should not be too easily dismissed. If it is possible to achieve a very low extinction risk per century, futures can easily become astronomical in size. We should not dismiss the probability of achieving very low extinction risk a priori.

Probable futures with astronomical amounts of value give rise to distinct problems with assessing the value of possible α-futures and our credences in them. In chapter 2, I will describe the EMV approach and discuss how large futures do not make it easier to apply EMV and might even make it more difficult. In chapter 3, I will look at the philosophical problem that comes from trying to give credences to possible α-futures. The very nature of these possible α-futures - some are radically different from the current world - makes it difficult to justify precise credences for these possible α-futures. Assessing their values and likelihood will require systematic and informed speculation, but it will remain speculation. Nonetheless, when speculation is all we have, I argue that it is sufficient to act on. Chapter 4 is more relevant to our actions. Rather than looking at the EMV of the future, I look at the EMV of extinction risk reduction. I describe some very positive side effects of reducing extinction risk, which could make the endeavour worthwhile even when the future looks negative.

Chapter 2. Problems with assessing the EMV of a possible α-future

2.1 A theoretical method to deal with moral uncertainty: expected moral value

2.1.1 Introducing expected moral value

The first issue we encounter in assessing the value of the future is how to assign a value to a single possible α-future when we are morally uncertain between multiple moral theories. In recent years the new field of moral uncertainty has produced a number of recommendations to deal with moral uncertainty. As could be expected, the field has no consensus (yet) on how to deal with moral uncertainty. In this thesis, I will apply one specific proposal: the *expected moral value approach*. This approach was developed by, among others, Sepielli (2013) and subsequently defended in MacAskill (2014).[15] Since moral uncertainty is most straightforward to interpret when moral propositions can be objectively true, I shall assume *moral cognitivism* (i.e. moral propositions can be true or false) for the rest of this thesis. However, before I explain EMV, let me first discuss the simplest proposal and how it fails to adequately deal with moral uncertainty. Seeing how the simple proposal fails sheds light on why EMV is a promising approach.

When faced with moral uncertainty, one intuition is to simply follow one's favourite moral theory: if the theory I most strongly believe is total utilitarianism, then I should just follow whatever it recommends and ignore all other theories because I have lower credences in them. Alternatively, we should choose the *option* most likely to be right (My Favourite Option). However, these approaches are mostly denounced by philosophers in this field. Both approaches only consider the *credences* of the decision-maker, but not the *values* of the theories or options. Consider the following:[16]

Jenny has 51% credence in theory T_1, and 49% credence in theory T_2. T_1 holds that option A is fantastic, and assigns it a value of 100, and finds option B horrific, assigning it a value of -100. In contrast, option A is horrific (-100) according to T_2, while it values option B a lot, and assigns it a value of 100. If these were the only two options, option A seems the moral choice, because Jenny has a slightly higher credence in T_1 than T_2. However, would things change if we introduce option C? Both theories believe option C to be worse than the

[15] MacAskill (2014) calls it *expected choiceworthiness*.

[16] Based on an example given in MacAskill (2014)

best option, but only by a little, and attribute a value of 99 to it. This makes the decision situation look as follows:

	T_1 - 51% credence	*T_2 - 49% credence*
A	100	-100
B	-100	100
C	99	99

Figure 1. A problem for My Favourite Theory and My Favourite Option.

Both My Favourite Theory and My Favourite Option recommend A, but surely option C is the best option! It minimizes the risk of doing something morally dejectable, and maximizes the probability of doing something morally commendable. Therefore, we need to take into account the values assigned to different options, and not only their credences.[17] The metanormative approach that does this employs *expected moral value* or *expected choiceworthiness*. Let me explain this approach.

It is generally accepted within philosophy and economics that when one faces empirical uncertainty about one's options, one should normally select the option with the highest expected value.[18] The expected value of an option is calculated by multiplying the probability of the possible outcomes by the value of these outcomes. Analogously, the EMV approach calculates the value of an option or outcome *A* by multiplying one's credence *Cr* in a theory T_i with the value $V_{T_i}(A)$ assigned to the option by theory T_i, such that

$$EMV(A) = \sum_{i=1}^{n} V_{T_i}(A) * Cr(T_i)$$

The EMV approach is elegant in its simplicity. It also captures the intuition that, besides credences, stakes matter as well. In the example case above, it assigns expected moral values of 2 to *A*, -2 to *B*, and 99 to *C;* this is in line with the notion that C was the far superior option. Furthermore, it is analogous to a well-established form of treating uncertainty: expected value has a solid theoretical basis (Briggs, 2017) and has fewer problems than other rational decision theories. Lastly, if one believes that moral

[17] For a more thorough criticism of these approaches, see MacAskill (2014).

[18] Exceptions include extremely small probabilities with extremely large payoffs (Bostrom, 2009), and maximizing expected value for one-shot decisions is somewhat controversial. This latter point seems relevant for this thesis, because (non-)extinction is modelled as a single choice - it is the ultimate one-shot decision. However, there is no obvious alternative besides extinction or non-extinction for risk-averse decision-makers.

uncertainty should be treated differently from empirical and other kinds of uncertainty, this seems to require decision-makers to know the nature of their uncertainty (MacAskill, 2014). However, sometimes it is unclear whether one's uncertainty stems from moral or nonmoral reasons. If one finds such uncertainty irrelevant to the decision procedure, this is another point in favour of EMV. Nonetheless, justifying the EMV approach is beyond the scope of this thesis and persuasive attempts have been made by others (cf. MacAskill (2014) for a case in favour of using EMV whenever possible, and Hedden (2016) for a criticism of it). Instead, I focus on the issues that arise when applying the EMV approach to extinction and the long-term future. This chapter will focus on the expected moral value of a single possible α-future, and examine the difficulties that arise. In chapter 3, I broaden my perspective to the EMV of the entire future.

2.1.2 The problem of intertheoretic value comparison

Although the EMV approach is attractive in its simplicity, it does have a number of requirements that are not obviously fulfilled. First, there is the requirement that theories can express a numerical value about an option (i.e. that they can be expressed on a *cardinal scale*). Many theories, such as rights-based theories or virtue ethics, do not do this by themselves, and often merely rank options. Although there is a method to translate ordinal preferences into a cardinal ranking,[19] some might resist such a method on theoretical grounds. This makes such theories harder to compare.

In the rest of this thesis I will focus on theories that can be represented on a cardinal scale, and therefore should be good candidates to apply the EMV approach too. This does not necessarily mean we can ignore what other theories say about the long-term future. Rather, I am limiting the scope of theories I consider to focus on the interesting and important problems that arise when one is uncertain between multiple theories with a cardinal scale. The less credence one puts in theories that cannot be represented on a cardinal scale, the more important these considerations will be for one's all-things-considered view.

However, even when two theories can be represented on a cardinal scale, there is a deeper issue: the problem of *intertheoretic value comparisons*. What is the meaning of phrases like 'the value of action A according to theory T_1 is twice as valuable as the value of action A according to T_2'? When theories are based on wholly different fundamental

[19] In brief, such a 'cardinalizing' method works as follows: an ordinal ranking can be transformed into a cardinal ranking if one can elicit ordinal preferences about probabilistic outcomes. For example, if a theory is indifferent between telling a lie and 0.1% chance of killing somebody, then killing somebody is a 1000 times worse than telling a lie according to that theory (Morgenstern & Von Neumann, 1953). Besides theories resisting such application, it is not always possible: some theories do not satisfy the axioms required.

assumptions are their values really comparable, or are they incommensurable? There is a wide range of possible answers here, ranging from ‘all theories are incommensurable’ to 'all theories are comparable on an intertheroretically valid cardinal scale’. In between, there is a pluralist view: some theories can be compared in a cardinal way (e.g. with the EMV approach), while others need some other way of comparison (cf. MacAskill, 2014 for such a pluralist view).

Before diving deeper into this issue, we will first look at a possible way around the problem. In the next section, I discuss such a possibility and argue that it fails. This means that the issue of intertheoretic value comparison comes back in full force, which I will explain in section 2.3.1.

2.2 Does Totalism dominate the EMV for large population cases?

2.2.1 In large population cases, unbounded moral theories overwhelm others

There are many reasonable moral theories one can have credence in and each theory will have something different to say about the value of an entire possible α-future. To incorporate this moral uncertainty, the EMV approach can be applied to scenarios regarding the long-term future. Because large population sizes are possible and probable in these scenarios, it is necessary to know how the implications of different theories compare for large populations.

Greaves & Ord (2017) considered how the EMV of a state of affairs (a possible α -future) would look like when the population size becomes very large. Because their argument is mathematical, they only consider mathematically well-behaving theories. They considered the following axiologies:[20]

Totalism: “the value of a state of affairs is the *sum* of the well-being of everyone in it—past, present, and future.”

Average View: “the value of a state of affairs is the *average* of the well-being of everyone in it—past, present, and future.”

[20] For simplicity, I leave out Critical Level views, which are a more general form of Totalism in which well-being is only valuable above a critical level, and has disvalue when it is below that level. What Greaves & Ord (2017) actually think to show is that Critical Level views are the view to defer to in large population cases, but this does not significantly alter the implications: critical level views are trivially easy to compare, so it would still resolve the moral uncertainty part of assessing the EMV of the long-term future.

Variable Views: creating extra persons has diminishing marginal value

Person-affecting views: only the well-being of presently or necessarily existing people matters morally.

Note that these are axiologies - they make statements about the value of possible α-futures, but not about what we *ought to do* (a deontic claim). For example, Totalism does not equal Total Utilitarianism, which adds the *deontic* claim that we must act to maximize the sum of everyone’s well-being.

Greaves & Ord (2017) show that, as population sizes become very large, most of these theories will differentiate less and less between the value of different possible α-futures. Take the Variable Views for example, the more people there are, the less the well-being of extra people will influence the value of a world. However, only one view, Totalism, does not have this property. Instead, as the population size in a possible α-future tends towards infinity, the value of that α-future will also tend towards infinity according to Totalism. In more general terms: Totalism is an *unbounded* view with respect to population size, while the other considered theories are bounded. As a result, the stakes of Totalism will ‘overwhelm’ the stakes of other theories in the EMV calculation when large populations are involved.

Interestingly, their approach sidesteps the problem of intertheoretic value comparison. They only look at the value functions of individual theories. If only one value function is unbounded with regards to population size, such a theory would dominate other theories no matter which (acceptable) cardinal method one uses to compare theories. The only requirement is that, in principle, the theories are comparable.

If Greaves & Ord (2017) are correct, this is either a *reductio ad absurdum* of the EMV approach or their analysis means that, when it comes to possible large populations, we should defer to whatever Totalism says because it is the only unbounded axiology. This would greatly help in assessing the EMV of the long-term future: moral uncertainty would no longer be an issue and the problem would be reduced to only empirical uncertainty - a nearly standard decision problem. Alas, as I show in the next section, it is not that simple.

2.2.2 Totalism is not the only unbounded view; enter Asymmetric Views

There is a population axiology that could rival Totalism and can resist being overwhelmed in the EMV calculation.[21] Here, I will first describe this family of views and what drives the

[21] This point was also raised recently by Plant & Kacmarek (Unpublished work).

arguments in favor of it. Then, I will discuss the issues that arise when there are competing moral views in large population cases.

Greaves and Ord note the criterion that such a theory needs to satisfy in order to resist being overwhelmed. They state the following (2017, p. 155):

> The theory must hold [...] not only that the alternative favored by the Total View is inferior in the large-population limit, but that the *amount* by which it is inferior grows without bound as the relevant population size increases.

I argue that Asymmetric Views fulfill this criterion and can therefore rival Totalism. They are a family of views that distinguish between the moral status of creating happy persons and the status of creating unhappy persons. Many people have the intuition that whereas creating happy persons is neither good nor bad, creating unhappy persons is definitely bad (Narveson, 1973). This is what has been called the Procreative Asymmetry, referring to the asymmetry between comparing net-positive lives and net-negative lives to non-existence (Mayerfeld, 1996; McMahan, 2009). I will refer to any view that weighs the creation of net-negative welfare stronger than the creation of net-positive welfare as an Asymmetric View. These views are often motivated by appealing to intuitions about intense suffering, such as the following unpleasant thought experiment.

Imagine the life of someone we will call Susan, who lives to the age of seventy-five. Although for large periods of her life she has a rather pleasant life, she has a horrible experience in the middle of her life. She is kidnapped by a psychopath and is submitted to extreme forms of physical and mental torture. She is gripped by a crushing and appropriate fear that it will never end and only get worse. After multiple weeks of intense suffering and complete desperation, she is suddenly released. After a long period of recovery, she lives out the rest of her life in a quiet a mildly pleasant way. The question is: if we knew this life story before it had happened, and if we had the choice to either bring Susan into existence or not, what should we have done? The asymmetry intuition is that her suffering is hard, or even impossible, to outweigh by the rest of her experience, and that it would have been better to not create Susan at all. Even more strongly driving this intuition, what if we had the option to create twenty people: one person who suffered immensely throughout their whole life and nineteen people leading happy lives. Can one person's suffering really be morally compensated for by other people's happiness?

There are many possible ways to incorporate this intuition. Because Asymmetric Views can potentially rival Totalism, we need to dig into the details, so let me give a quick overview of the possibilities of incorporating the asymmetry intuition (cf. Arrhenius et al. (1994) for a comprehensive overview).

Some views give priority to any amount of suffering over any amount of happiness[22], these are Strict or Lexical Asymmetric Views. These latter two Asymmetric Views will *always* regard the expected value of creating a life as negative, because there is always a possibility that the life has negative welfare and if it has positive welfare the utility is zero. In addition, this would imply that the survival of humanity is always negative if only welfare is intrinsically valuable. Although for most people these implications are counterintuitive enough to discredit such views, some decide to bite the bullet (cf. Benatar, 2008). However, these views will *always* be in favour of extinction over any population size, so it is also not an *interesting* rival to Totalism. Because of the implausibility and theoretical uninterestingness of Strict and Lexical views, from here onward I will consider only Moderate Asymmetric Views: those that weigh suffering more heavily than happiness when aggregating the total value of a world or act, but still weigh happiness or positive well-being to some extent. Nonetheless, we shall see that including these views makes assessing the expected moral value of a possible α-future troublesome.

Besides being strict or moderate, Asymmetric Views can also be personal or impersonal. Personal Asymmetric Views draw a distinction between the creation of net-positive and net-negative *lives*. On these views, creating a net-negative life is worse than creating a net-positive life is good, but suffering and happiness within a life receive equal moral weight. In contrast, Impersonal Asymmetric Views instead draw a distinction between net-positive and net-negative *welfare moments* (Arrhenius et al., 1994). On these views, a moment of suffering receives more moral weight than an equally strong moment of happiness, regardless of whether the moments occur in the same person or in different persons. I will not take sides on which distinction should be preferred, and interested readers may consult - among others - Mayerfeld (1996), or Arrhenius et al. (1994). When drawing a distinction between welfare moments, one needs to defend that it is meaningful to state something like ‘suffering is worse than happiness is good’. When drawing the distinction between lives, one needs to defend why individuals are the relevant unit of analysis. Some authors also dispute that happiness and suffering can be compared objectively (Knutsson, 2016), with the result that such comparisons are either subjective or meaningless. Nonetheless, Knutsson still maintains we should prioritize preventing suffering over promoting happiness.

[22] ‘Happiness’ is a slightly ambiguous term, as it is sometimes understood as hedonistic happiness only: pleasurable experiences. I use it in a broader sense: as the opposite of suffering. I therefore also use (net-)positive well-being and happiness interchangeably.

From here onwards, I shall discuss the asymmetry as if it only pertains to comparing the creation of *lives* and not welfare moments. I use ‘net-positive (or negative) lives’ as shorthand for ‘lives with, in aggregate, a positive (or negative) well-being. When I write that “a net-positive life *weighs* as much as (or weighs less than) a net-negative life” I mean that “creating a life that has, all things considered, a positive well-being *is* just as good as (or worse than) creating a life that has, all things considered, a negative well-being of the same magnitude (but in the other direction), such that creating both lives in parallel is morally neutral (or morally negative).”

My focus on lives, rather than welfare moments, allows me to keep things simple, and it should not significantly alter my conclusions if one replaces lives with welfare moments. Furthermore, I shall assume that all lives have a quality that can be compared meaningfully. Lastly, I ignore views that weigh the welfare of current beings more than the welfare of future beings (presentist-leaning views), because I am interested in the expected moral value of creating future beings *if* they matter. For the sake of simplicity, I also ignore views that assign variable moral weight to welfare such that the alleviation of suffering increases in importance the more suffering there is, and/or such that the promotion of happiness decreases in importance the more happiness there is (e.g. Prioritarianism, cf. Parfit, 2012).[23]

To show the significance of slightly different beliefs about moral parameters and to show how different theories ‘compete’ against each other, I will introduce some formal notation. The relative weights attributed to the moral value of creating net-positive lives and net-negative lives can be framed by a ratio.[24] Let the N-ratio (for “normative ratio”) $N(T_i)$ of a theory T_i be:

$$N(T_i) = \frac{moral\ value\ of\ creating\ a\ life\ with\ net{-}positive\ well{-}being}{moral\ value\ of\ creating\ a\ life\ with\ net{-}negative\ well{-}being}$$

For Totalism the N-ratio is $\frac{1}{1}$; for Strict Asymmetrism it is $\frac{0}{1}$; the N-ratio of Moderate Asymmetric Views (MAV’s) lies in between: $\frac{0}{1} < N(MAV) < \frac{1}{1}$. As we shall see, including MAV’s severely complicates the application of EMV to decision situations which involve large population sizes. To compare what the different theories say in large population

[23] Prioritarianism might in fact be a rival to Totalism in large population limits: although it will care less and less about increasing positive well-being, it will care more and more about preventing negative well-being, because in larger populations there will be more, and probably more intense, negative well-being. Prioritarianism will have a function that is unbounded in the negative direction, and it may or may not grow faster than Totalism’s function.

[24] This ratio is adapted from Althaus (2018). Note that Althaus refers to creating goods (value) and bads (disvalue) in general, while I restrict it to whole lives only.

cases, we need to know their value functions. The value function looks as follows for a possible α-future F_j:

$$\textit{Totalism:} \qquad V(F_j) = \overline{w(F_j)}\, q(F_j)$$

$\overline{w(F_j)}$ represents the average well-being of the total population in possible α-future F_j, and $q(F_j)$ represents the total population that lives in the possible α-future F_j.

The value function of a Moderate Asymmetric View looks more complicated. Whereas the views considered by Greaves and Ord only care about the average level of well-being among the total population, MAV's need to distinguish between the average well-being of presently existing persons $\overline{w(P_j)}$, average well-being of future persons with net-positive well-being $\overline{w^+(F_j)}$, and the average well-being of future persons with net-negative well-being $\overline{w^-(F_j)}$. Note that it is controversial when a net-negative life becomes net-positive in terms of well-being. However, I merely require the assumption that affective states can be compared between persons, and states that feel positive (e.g. amusement, inspiration) contribute positively towards overall well-being, while states that feel negative (e.g. grief, depression) contribute negatively towards overall well-being. A net-neutral overall well-being is therefore equivalent to no well-being at all, or to not feeling anything ever (which is only a theoretical possibility for humans). Besides information on average well-being levels, MAV's also need information on the *amount* of happy and unhappy lives.

If we know these parameters, we can define a vector $\vec{v}\,(N,F)$ which represents the average value of an 'extra' person. If we can define this vector, we end up with a very simple multiplication formula for the value of the α-future which contains one scalar (population size) and one vector (the average value of an extra person). This vector will be important, because it will determine the moral value of a future population. If adding an extra person to a population has, on average, a negative moral value, then surely we should not want a large population! However, as we shall see, the N-ratios need to be normalized before vectors can be compared between different theories. This is the, admittedly inelegant, formula for calculating the vector:

$$\vec{v}\,(N,F_j) = \frac{N * \overline{w^+(F_j)} * q^+(F_j) + \overline{w^-(F_j)} * q^-(F_j)}{q(F_j)}$$

This formula looks quite complicated, but it is actually quite simple. The first part is $N * \overline{w^{+}(F_j)} * q^{+}(F_j)$. This is the value of the part of the population that enjoys lives of positive well-being, and it is 'discounted' by the N-ratio (the relative weight of creating net-positive compared to creating net-negative lives). The second part, $\overline{w^{-}(F_j)} * q^{-}(F_j)$, represents the part of the population that suffers lives of negative well-being. This is always a negative number (because the population is never negative, and the average well-being is by definition negative). The third part is '*divided by* $q(F_j)$'. This gives us the average moral value per person. This last part seems unnecessary, but it helps us to easily represent the value of a future population by the formula below.[25] Note that when $N = \frac{1}{1}$, $\vec{v} = \overline{w(F_j)}$, meaning that this function can also represent Totalism. Now we can formalize the value $V(F_j)$ of a possible α-future F_j according to Asymmetrism in the following ways:

Asymmetrism: $V(N, F_j) = \vec{v}\ q(F_j)$ or

$$V(N, F_j) = N * \overline{w^{+}(F_j)} * q^{+}(F_j) + \overline{w^{-}(F_j)} * q^{-}(F_j)$$

This function will grow unboundedly with $q(F_j)$, in either the positive or negative direction, depending on the sign of the vector $\vec{v}$. As $q(F_j) \to \infty$, $V(F_j) \to \infty$ or $-\infty$, unless $\vec{v}$ equals *exactly* zero which is unlikely. Furthermore, this function *can* outgrow Totalism if $|\vec{v}\ | > |\overline{w(F_j)}\,|$, and thus fulfill the criterion set by Greaves and Ord mentioned earlier. This implies that it is no longer certain that the effective axiology defers to the preference of Totalism in cases of very large populations: it will depend on the N-ratios and the values of the empirical parameters. Furthermore, comparing vectors requires intertheoretical value comparison. This means that moral uncertainty can, unfortunately, not be sidestepped. Instead, we need to deal with moral uncertainty, which brings us back to one familiar problem (intertheoretic value comparison, section 2.3.1) and one possible new problem (disagreement about large values, section 2.3.2).

2.3 Can the EMV approach resolve (big) disagreement?

2.3.1 Disagreement and the problem of intertheoretic value comparison

Given that we have to compare multiple moral theories that have something significant to say about the value of the long-term future, we should find out under which conditions

[25] For the sake of simplicity, we leave present beings out of the picture.

these theories disagree. I will first set out some examples to find out where Totalism and Asymmetric Views disagree.

Suppose we have credence in two theories. On the one side we have Totalism, according to which the value of a possible α-future is the sum of the well-being of all individual lives in that α-future. The second theory is a Moderate Asymmetric View with an N-ratio of $\frac{1}{10}$ ($\mathrm{MAV}_{0.1}$)[26], according to which the creation of lives with net-negative well-being weighs ten times as heavy in the value function as the creation of lives with net-positive well-being, if the levels of well-being are of the same magnitude (i.e. if $\overline{w^{+}(F_j)} = -\overline{w^{-}(F_j)}$). In addition, let us consider the following trajectories to highlight some disagreements. For each trajectory, I denote the amount of individuals with net-positive lives and those with net-negative lives, and their respective average well-being levels. The two right-hand columns represent the value attributed to the trajectory.

Trajectory	**Positive well-being** $q^{+}(F_j) * \overline{w^{+}(F_j)}$	**Negative well-being** $q^{-}(F_j) * \overline{w^{-}(F_j)}$	**Value (Totalism,** $\boldsymbol{N = \frac{1}{1}}$ **)**	**Value** ($\mathbf{MAV}_{0.1}$**,** $\boldsymbol{N = \frac{1}{10}}$ **)**
Utopia	$10^{21} * 10^{3}$	$10^{15} * -(10^{3})$	$\approx 10^{24}$	$\approx 10^{23}$
Decent & Big	$10^{21} * 10^{3}$	$.5 * 10^{21} * -(10^{3})$	$0.5 * 10^{24}$	$-0.4 * 10^{24}$
Dystopia	$10^{15} * 10^{3}$	$10^{21} * -(10^{3})$	$\approx -10^{24}$	$\approx -10^{24}$
Flawed realization	$10^{21} * 10^{-3}$	$10^{21} * -(10)^{-3}$	0	$-9 * 10^{17}$
Plateau	$10^{16} * 10^{2}$	$10^{15} * -(10^{2})$	$9 * 10^{17}$	0
Instant Extinction	*0 * 0*	*0 * 0*	0	0

Figure 3. Expected Value according to Totalism and Moderate Asymmetric View which disvalues the creation of net-negative lives 10 times as heavy as net-positive lives ($\mathrm{MAV}_{0.1}$). Note also that the values according to the two theories are not yet intertheoretically comparable, and that the values do not represent actual estimates.

[26] I use the notation MAV_{N} where N represents the ratio of weights between creating net-positive and net-negative lives.

Given moral uncertainty between the above two theories, which α-futures are worse than extinction? And which are better? To answer this question, we can first invoke a principle of weak dominance:

> *Principle of Weak Dominance:* when no moral theory judges α-future *X* worse than α-future *Y*, and at least one theory judges α-future *X* to be better than α-future *B*, α-future *X* has a higher expected moral value than α-future *Y*.

With this principle, we see that Utopia strongly dominates Instant Extinction and Plateau weakly dominates Instant Extinction; these α-futures are more desirable than Instant Extinction. Furthermore, Instant Extinction weakly dominates Flawed Realization, and strongly dominates Dystopia; these α-futures are less desirable than Instant Extinction. However, Decent & Big trajectories are controversial: worlds in which there are more happy lives than unhappy lives, but not enough happy lives that even the Asymmetric view holds that it is positive.

In general, these particular theories would disagree about whether possible α -futures have a higher or lower EMV than instant Extinction. An α-future lies in the *disagreement zone*[27] when the amount of positive well-being is larger than the amount of negative well-being (so that Totalism would still be in favour of creating it) but less than ten times as large (so that Asymmetrism would still be against creating it). Figure 4 illustrates this disagreement zone for the considered two moral theories.

[27] Note that this point was raised as well in a manuscript by Plant & Kaczmarek (2018). They call it the 'disagreement danger zone'.

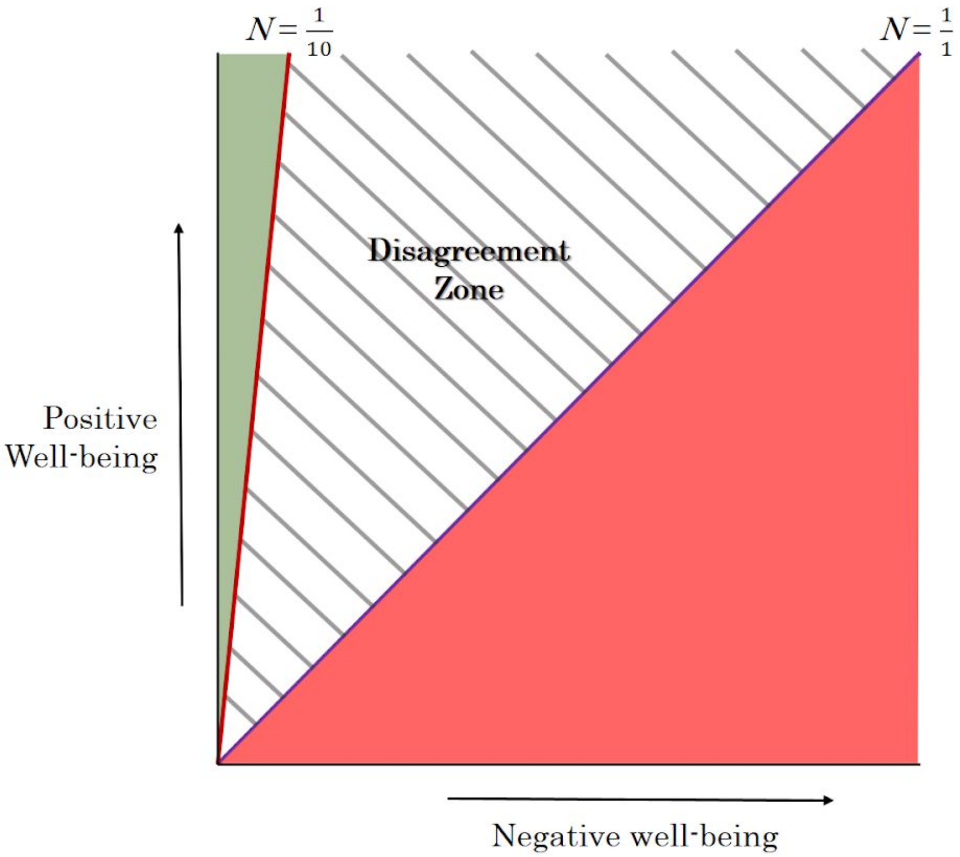


Figure 4. Example of a disagreement zone for two theories, one with an N-ratio of $\frac{1}{1}$ (like Totalism), another with an N-ratio of $\frac{1}{10}$. In the disagreement zone, Totalism will value the world as positive, while Asymmetrism would value it as negative. The green area indicates possible α-futures in which both theories agree that the α-futures have positive value, and the red zone indicates trajectories in which both theories agree that the trajectories have negative value.

When the Principle of Weak Dominance is not decisive, the possible α-future lies in the disagreement zone. A possible α-future falls in the disagreement zone when the E-ratio (i.e. $\frac{\textit{total well-being of net-positive lives}}{\textit{absolute total well-being of net-negative lives}}$) is larger than one (i.e. there is more positive well-being than negative well-being), yet smaller than one divided be the smallest *N:*[28]

$$\frac{1}{N_{max}} < E - ratio < \frac{1}{N_{min}}$$

[28] Note that N is, in any reasonable case, smaller than or equal to one. It would be unreasonable to value creating happiness *more* than creating suffering. If E is equal to or larger than 1/N, that means $E * N$ will be equal to or larger than 1. When that is the case, there is more value than disvalue according to the theory with that particular N-ratio and according to theories with a smaller N-ratio.

For those trajectories, it will be necessary to calculate the expected moral value of a trajectory more precisely. Although there would be disagreement, the EMV approach can still decide when there is disagreement, because *the stakes of the disagreement matter too.* In fact, when only Totalism and a finite number of MAV's are considered, the EMV formula for a particular possible α-future can be represented by the population size times a credence-weighted vector:

$$EMV(F_j) = q(F_j) \sum_{i=1}^{n} \vec{v}_{T_i} * Cr(T_i)$$

In this formula, F_j refers to a single possible α-future, and *Cr(*T_i*)* to one's credence in theory T_i. This formula would yield a positive value if the credence-weighted vector is positive, and negative vice versa. This formula is also equivalent to $\sum_{i=i}^{n} V_{T_i}(F_j) * Cr(T_i)$, in which $V_{T_i}(F_j)$ represents the value of possible α-future F_j according to theory T_i. However, there are two potential problems with this approach. The first problem is the problem of *intertheoretic value comparison.* Recall the formula I used earlier:

$$\vec{v}\,(N, F_j) = \frac{N * \overline{w^{+}(F_j)} * q^{+}(F_j) + \overline{w^{-}(F_j)} * q^{-}(F_j)}{q(F_j)}$$

Although it appears that we have all the information we need, $\vec{v}_{T_i}$ contains a term which is not easily comparable across different theories: *N,* which represent the relative weight of the value of creating a net-positive life vs. the disvalue of creating a net-negative life. The challenge of intertheoretic value comparison is this: how should we compare $N = \frac{1}{1}$ to $N = \frac{1}{10}$? There are at least two options.

First, it could be that the MAV (Moderate Asymmetric View) values creating net-negative lives the same way Totalism does, but the MAV puts comparatively less weight on the creation of net-positive lives. So, for example, on both views, creating a net-negative life is worth -1, but creating a net-positive life is worth +1 for Totalism and only +1/10 for the MAV (see perspective A in the figure below).

However, it could also be that the MAV values creating net-positive lives the same way that Totalism does, but that the MAV puts relatively *more* weight on the creation of net-negative lives. So, for example, creating a net-positive life is worth +1 on both views, but creating a net-negative life is worth -1 for Totalism and -10 for the MAV (see perspective B in the figure below).

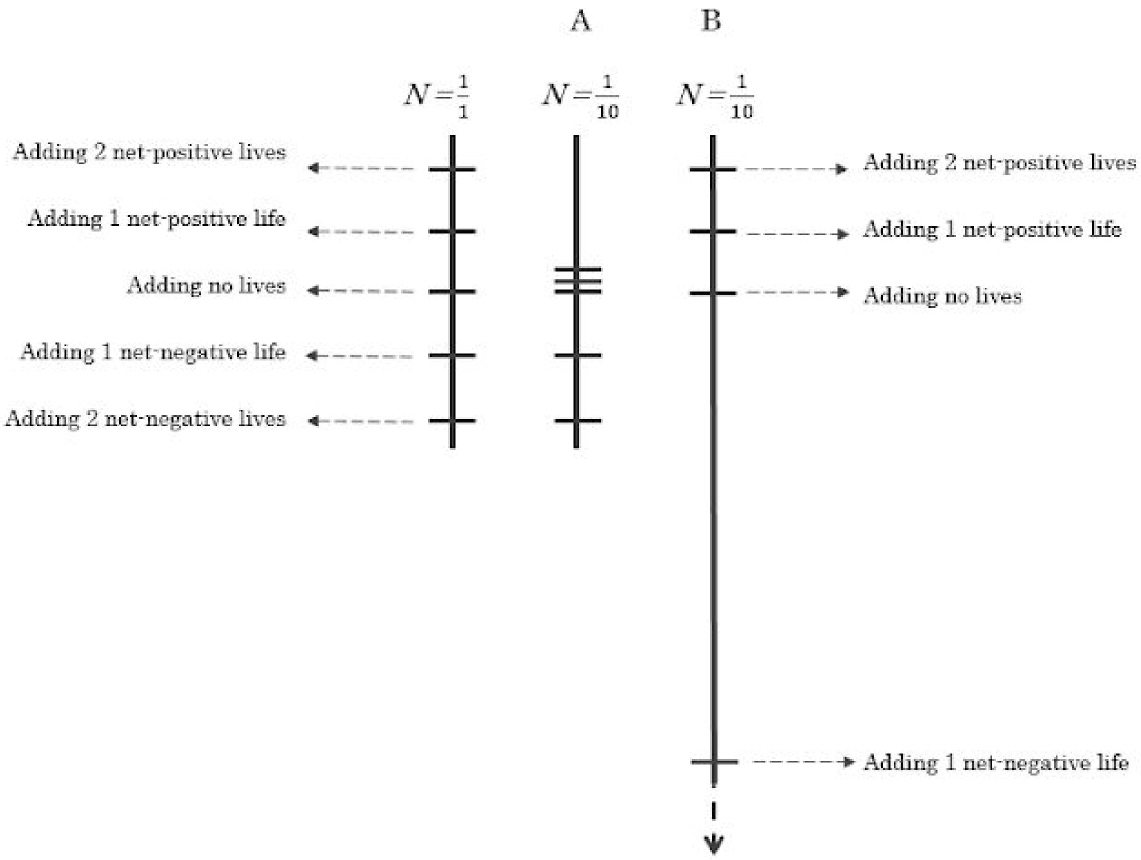

A
B
$N=\frac{1}{1}$
$N=\frac{1}{10}$
$N=\frac{1}{10}$
Adding 2 net-positive lives
Adding 1 net-positive life
Adding no lives
Adding 1 net-negative life
Adding 2 net-negative lives
Adding 2 net-positive lives
Adding 1 net-positive life
Adding no lives
Adding 1 net-negative life

Figure 5. Different ways of comparing Totalism to a Moderate Asymmetric View which weighs the creating of net-negative lives ten times heavier than creating net-positive lives. Is A or B the correct way to compare the theories, or should we choose another option?

Ideally, we know what value each theory assigns to a possible α-future on a cardinal scale (i.e. assigning a numerical value to every option) that is intertheoretically valid. However, how could we find this intertheoretically valid scale? Let me demonstrate the difficulty.

We could calculate a single vector $\vec{v}$ by taking the credence-weighted average normative ratio $\overline{N}$. But how should the credence-weighted average normative ratio $\overline{N}$ be calculated? Suppose our credences are equally split between an N-ratio of $\frac{1}{1}$ and an N-ratio of $\frac{1}{10}$. How to resolve this depends on which claim we believe Asymmetric Views to be making: are they weighing creating net-positive lives *less* than Totalism (perspective A), or weighing creating net-negative lives *more* than Totalism (perspective B)? If it's the first perspective, it seems we should normalize N as follows: $\overline{N} = .5 * \frac{1}{1} + .5 * \frac{1}{10} = \frac{5.5}{10} = 0.55$.

However, in such a case the average will always be much closer (in fractional terms) to Totalism's weight than Asymmetrism's weight. The average weight will always be dominated by the relatively high value of $\frac{1}{1}$, and when one has a credence of 50% in Totalism, the average N will never be less than $\frac{1}{2}$. For example, even if Asymmetrism barely weighs the creation of net-positive lives, e.g. $N = \frac{1}{100}$, the average still is 0.505, which nearly equals $\frac{1}{2}$. This seems unfair favoritism of Totalism over other theories. If Asymmetrism weighs creating net-negative lives *more* than Totalism (perspective B), we should use a different calculation that gives more weight to Asymmetric Views. For example, $\overline{N} = \frac{1}{(1+10)/2} = \frac{1}{5.5} \approx .182$. This seems conceptually more accurate in representing our uncertainty about moral weights. In conclusion, it is unclear how to take the normalized N-ratio over all theories, but the average does not seem the right way.

Another possibility is that Asymmetrism is silent on how it compares to Totalism, and only claims that in its own theory creating net-positive lives weighs less than creating net-negative lives.[29] In that case, we probably need to turn to structure-based normalization methods, rather than content-based ones. The most promising structure-based method seems to be *variance normalization.* This method makes different moral theories comparable by rescaling them so that they have the same variance. Metaphorically, one can view this as assigning each theory an equal amount of 'credit' and each theory assigns this credit differently over the possible options (MacAskill, 2014).

Variance normalization has a number of advantages over other structure-based methods of normalization, as described in MacAskill (2014). Among others, it can account for unbounded theories, which cannot be normalized by their range, minimum, or maximum because unbounded theories do not have these.[30] Although variance normalization is elegant in theory, it is difficult to apply in practice. The method requires knowing the variance of all considered theories, given a set of options. However, if we care about *all available options* we encounter two problems. First, we do not know the range of all available options (in terms of population size $q(F_j)$ and other parameters). Second, even if we could specify the range, we would not know how to distribute the probabilities over these possible α-futures. The different normative theories' variances change differently depending on how credences are distributed over the option space. A different credence

[29] The benefit of using ratios is that it is silent on how to compare the different value assessments between theories.

[30] Of our considered theories (Totalism, (Im)personal Strict Asymmetrism, (Im)personal Moderate Asymmetrism), only (Im)personal Strict Asymmetric Views have one of these: a maximum where no beings (on the personal perspective) or disvalue (on the impersonal perspective) exists.

distribution over options gives different variances. Therefore it is practically very difficult to non-arbitrarily normalize two or more theories by their variance.

In conclusion, we should expect a number of possible α-futures to be in a *disagreement zone;* some theories will prefer such α-futures over extinction, while other theories will prefer it the other way around. This brings us to the problem of intertheoretic value comparison, which is not a fully resolved philosophical problem and hard to do in practice. However, the problems in assessing the EMV of the entire long-term future do not stop here. In the next section, I discuss an additional concern that one might have in cases of moral uncertainty and large populations.

2.3.2 Big worlds breed big disagreement

The EMV approach is a good way to take relative stakes of different theories into account. If credences are equally distributed between two different theories, the theory that holds that there is more at stake gets to decide. However, one might believe that we should not apply EMV in large population cases. In this section, I discuss why people might believe that and consequently argue why it would be mistaken to believe that.

Let's return to the disagreement zone. In this zone, some theories will scale the value of the possible α-future by a positive vector, and some by a negative vector. Scaling such worlds by a negative, as Asymmetrism would hold is right, results in an astronomically large negative number, akin to dystopia! Scaling such worlds by a positive, as Totalism would hold is right, resulting in an astronomically large positive number, akin to utopia! And not creating the world would, according to Totalism, be as worse as creating dystopia! Thus we see: *big worlds breed big disagreement.* As the disagreement moves further into the disagreement zone, the relative stakes become smaller. Given such enormous disagreement, I believe many readers would wonder if this is a different form of disagreement which cannot be solved quantitatively. Let's discuss the following proposition:

> *Deep Disagreement Intuition:* when the difference in value attributed by different theories to a possible α-future becomes significantly large (in comparison to the values in daily decision contexts), moral uncertainty between those theories should not be handled by the expected moral value approach.

The 'should' in this proposition refers to what is the correct or rational way to deal with moral uncertainty (cf. 'the ought of moral uncertainty' by Bykvist, 2017). There are two ways of motivating this deep disagreement. First, the intuition might be driven by *large values* (relative to daily decision contexts): at some scale a certain theory holds there is so much at stake that it can no longer be overwhelmed by another. Second, the intuition might be driven by *relatively close values*: at some point, the values attributed by different theories to a possible α-future lie so close to each other (e.g. within one percentage), that it would be overzealous to still apply the expected moral value approach.

Against the intuition of deep disagreement, I pose the following proposition:

> *Resolvable Disagreement Proposition:* if different theories both can be represented by a fully cardinal structure and can be intertheoretically compared, moral uncertainty can always be handled by the expected moral value approach, regardless of the scale of the problem or the closeness of the values attributed by different theories

I believe neither motivation for the Deep Disagreement Intuition is justified. Let's discuss the first motivation for deep disagreement, that "at some threshold a certain theory holds there is so much at stake (relative to daily decision contexts) that it can no longer be overwhelmed by another." For example, it could state that 'when a theory holds there is a million times more at stake than the death of a single individual, that theory can no longer be overwhelmed by another.' The idea behind this is that large amounts of value cannot simply be 'brushed aside' and that it feels like the EMV approach does precisely that; applying a simple calculation just feels wrong. To start, we should note that the theory should be *credence-weighted* (i.e. multiplied by the subjective degree of belief in the theory). If it were not, any possible but highly improbable view could easily claim deep disagreement and stagnate every decision procedure.

However, this intuition faces at least the following two problems. First, it is unclear where the threshold should be set because any 'daily decision context' is an arbitrary reference point and any scale is also arbitrary. Second, this intuition is suspiciously similar to the cognitive bias *scope neglect* (Baron & Greene, 1996). Imagine there are only two options: action A and action B. Assume that theories T_1 and T_2 are completely comparable. Theory T_1 holds that doing action A is as bad as the death of 10^{23} people and action B is neutral, while theory T_2 holds that doing action B is as bad as the death of 10^{24} people and action A is neutral. One might say that, because there are so many lives involved, neither option is better than the other. But choosing action B is as bad as condemning 900,000,000,000,000,000,000,000 *more* people to death. If you believe 'high stakes cannot be ignored', these people should be taken into account. Human brains have not evolved to deal with large sizes and consequently represent different large sizes conceptually as 'similarly large' even when they differ by multiple orders of magnitude, which (unsurprisingly) leads to irrational decisions. Unless there is a strong theoretical justification for a threshold at a particular scale, this intuition seems to be a cognitive bias and we should *not* adjust our metanormative principles to biases.

The second way of motivating the Deep Disagreement Intuition also does not hold up to scrutiny. It stated that "at some point, the values attributed by different theories to a possible α-future lie so close to each other (e.g. within one percentage point), that it would be overzealous to still apply the expected moral value approach." However, this is - again - ignoring that we are talking about really large values (relative to daily decision contexts). Take the earlier example, but instead theory T_1 holds that doing action A is as bad as the death of $9.91 * 10^{25}$ people and action B is neutral, while T_2 holds that action B is as bad as the death of $10.00 * 10^{25}$ people and action A is neutral. The difference in value is less than a percentage, but that does not mean the difference should not be taken seriously! Choosing action B is still as bad as condemning 900,000,000,000,000,000,000,000 *more* people to death.

Therefore, I maintain that the EMV approach should be applied at any scale, at least when theories are intertheoretically comparable. However, I believe the second intuition points towards another potential reason to not apply the EMV approach: when the relative stakes are so close together it implies that our judgments are based on very fragile arguments. Slightly more information and deliberation could alter the balance. This point is better expressed differently than by deep disagreement. Furthermore, it arises most commonly from empirical, not moral, uncertainty. Empirical uncertainty enters the stage as we move away from assessing the EMV of single possible α-futures towards assessing the EMV of the entire long-term future. In the next chapter, I discuss *fragile EMV*. Although it is a stronger reason to reject applying the EMV approach to large population cases, I maintain that it is not strong enough.

Chapter 3. Problems with assessing the EMV of the entire long-term future

3.1 Assessing the EMV of the entire long-term future introduces empirical uncertainty

If we leave the problems with assessing the expected moral value of single trajectories aside for a moment, we can bring further problems to light. If we want to assess the EMV of the *entire* long-term future, we need to handle not only moral uncertainty, but also empirical uncertainty. Remember the formula from chapter 1:

$$EMV(F) = \sum_{j=1}^{m} \sum_{i=1}^{n} V_{T_i}(F_j) * Cr(T_i) * Cr(F_j)$$

To utilize this formula, we need to assign credences to which α-future will actualize.

This would be difficult for any timespan, but the fact that we are talking about the *long-term future* makes this extra difficult. As mentioned in chapter 1, we will need to assess the likelihood of very large α-futures and that requires engaging with futuristic considerations and informed, systematic speculation. For the sake of illustrating what such considerations and speculations look like, allow me to briefly speculate on two issues relevant to the EMV of the long-term futures: how futures can come about that are astronomically worse than an empty universe, and how they weigh up against the probability of α-futures that are astronomically better than an empty universe.

First, several scholars (Sotala & Gloor, 2017; Torres, 2018) have raised the possibility of *suffering risks*: the risk of "events that would bring about suffering on an astronomical scale" (Althaus & Gloor, 2016).[31] Sotala and Gloor (2017) list a number of possible ways astronomical amounts of suffering could come about. For example, the dominant agents may cause suffering simply by not prioritizing the well-being of other beings in a similar way many animals currently suffer in factory farms. They suffer not because we want to hurt them, but because a situation emerged in which people and organizations prioritize something else (convenience, profit) over animal welfare. Additionally, if it ever becomes possible to simulate or emulate sentient beings - which should not be ruled out, given the timescales we are considering - vast amounts of sentient beings might be created. It then

[31] Note that Bostrom (2003) does reserve the term *hyperexistential risk* to indicate scenarios with large scope (transgenerational and cosmic) combined with a 'Hellish' quality.

becomes important for which reasons these beings would be created: for altruistic reasons, economic reasons, or other selfish reasons? If sentient beings are created for economic reasons, their existence will be optimized for economic productivity, not well-being. It is unclear whether such a life would be worth living (cf. Bostrom (2014b) or Hanson (2016) for some discussion of negative well-being in futuristic scenarios). However, sentience might in some cases even be optimized for suffering. Although this seems relatively unlikely compared to optimization for positive value, the possibility cannot be dismissed out of hand. To adequately assess the EMV of the long-term future, one has to speculate about scenarios with unfamiliar types of actors, with unfamiliar types of minds, and unfamiliar types of motivations.

These speculations are necessary to assess the EMV of the entire long-term future. Now, this makes the EMV very speculative, but is it *too* speculative? In the next section, I make this consideration more precise and briefly survey the options we have to deal with it.

3.2 Strong empirical uncertainty makes the EMV fragile

What is the problem when something is *too* speculative? To illustrate what might be wrong with acting on speculation and to provoke some intuitions, let’s consider the following thought experiment.

> *Horror Facility.* Suppose you have a friend, Tom, who is terminally ill. You receive the following phone call: ‘Hello. Tom has enlisted in our medical facility for treatment. Beware that we are no ordinary facility: we can cure any disease and guarantee a healthy and happy life to anyone we choose to treat. However, to achieve and improve our skill, we use half of our patients as test subjects. Tom is currently in a coma, and you have to decide in an hour to which ward he will be transported. Goodbye.’ Frantically, you start doing research. You find out the following. There are two wards, Ward A and Ward B; one of them is where patients are treated, and one where they are experimented with. The claim about treatment looks sound: people really have been cured. Furthermore, you discovered that the experimentation facility is absolutely horrific: people are tortured and kept in horrible conditions for the rest of their lives. You also found out one tiny detail: Amanda, one of the staff members of Ward A, is vegetarian. You look at the clock; time’s up. You know for a fact that if you had had more time, you could have found out which ward is the treatment ward.

What do you choose for your friend Tom? What are you rationally and morally required to choose? It seems permissible to choose either ward. Sure, you have *some*, albeit very flimsy, evidence that Ward A is better than Ward B. The fact that Amanda is vegetarian *might* indicate that she cares about morals, and that *might* indicate that she doesn't condone experimenting on humans and thus works in the treatment ward rather than the experimentation ward. However, it seems ridiculous to be required to make a choice based on such flimsy evidence, especially if more evidence is available: the choice is not always random. In the rest of this section I will argue that we are in a comparable state regarding the long-term future of humanity: our evidence is flimsy, we cannot consult with the people we are making the choice for, and the wrong choice will be catastrophic. To illustrate the problem, take the following thought experiment:

> *Extinction Button.* Suppose you are in the disagreement zone, such that you have some credence in theories which imply that the future is negative in expectation and some credence in theories which imply that it is positive. Now suppose that there is a button in front of you. This button gives you the option to permanently end all sentient life in the universe if you press it. After one hour, the button disappears forever. If you do not press it, only you will die. It will be a painless death, but we need to ensure that your choice is not biased by egoistic motivations, nor that you can influence the trajectory the world is on. Although you are highly uncertain, you believe the EMV of the entire long-term future is positive on the fragile basis that happiness and suffering are equally energy-efficient, but future agents are more likely to optimize for happiness than for suffering. You decide to not end all sentient life permanently. Thirty years later, scientists find a theoretical proof that suffering is actually *much* more energy-efficient to create than happiness. Instead of having created value as good as creating Utopia, you have made an incredibly harmful decision akin to creating Dystopia.

This is what could happen if our EMV is based on highly uncertain predictions about the long-term future, and on highly uncertain credences in moral theories that weigh happiness and suffering differently. We see that, as population size increases, the EMV becomes increasingly sensitive to the empirical parameters and one's credences in different moral theories. These are all parameters that we should be extremely uncertain about. Even if our best guesses for these parameters result in a highly positive EMV, should we really act on such flimsy evidence? A very slight change to our set of evidence could radically change the EMV, changing it from highly positive to highly negative.

To make this more concrete, we can borrow terminology from Joyce (2005), who distinguishes between the *balance, weight,* and *specificity*[32] of evidence. Joyce describes the balance as "how decisively the data tells for or against the hypothesis" and the weight as "the gross amount of data available" (Joyce, 2005, p. 158). For example, consider an urn with 3 balls, of which *i* balls are blue, and *3 minus i* balls are green. If you see three draws (with replacement after each draw) and two out of three balls are blue, your credence is highest for the hypothesis that *i = 2* (that the urn contains 2 blue balls). However, if you see an additional 27 draws, with a total of 17 blue balls and 10 green balls, you still believe *i = 2* is more likely, but to a much higher extent. More importantly, you could draw another 30 balls, but that would be highly unlikely to change your credence significantly. The second 30 draws have a much lower *information value* than the first 30 draws. The point to demonstrate is that our credences reflect the balance of evidence, but they do not reflect how strongly our credences should *respond to new evidence.*

In our case, the balance of evidence is reflected in the *point estimates* (a single number, rather than an interval) of the different variables. Our EMV calculation is entirely based on the balance of evidence (whatever it may be), and thus on point estimates. Whereas the balance is reflected in our EMV, the weight of evidence behind it is not.

Fragile credences are the opposite of resilient credences (Joyce, 2005). A credence is fragile if its unconditional probability is very different from its probability conditional on an extra piece of evidence. Thus, when the weight of evidence is light, credences are fragile. In the example your credences are fragile about how many blue balls the urn contains after one draw, because they can change a lot when new information comes in. However, they are quite resilient after thirty draws. Fragile credences can 'infect' our expected value calculation. Let me define *fragile expected moral value* as 'expected moral value which contains fragile credences that can easily flip the sign of the value when new information comes available.'

However, not all evidence weighs equally, some evidence matters more than other. Bostrom (2014a) coined the concept of *crucial considerations*: "[considerations] that radically change the expected value of pursuing some high-level subgoal." In this case, the high-level subgoal is preventing extinction. The expected value of pursuing that goal changes radically if we acquire new information that implies the expected value of the future is negative. However, there can be multiple pieces of information, each switching

[32] For simplicity's sake we leave out specificity as it does not affect the gist of the argument.

the sign of expected value from positive to negative, or vice versa. How should one act if it is likely that there is at least one crucial consideration out there, if it is uncertain how easy it is to uncover the crucial consideration(s), and if it is uncertain *how many* crucial considerations are out there? Fragile credences and crucial considerations imply that the value of more information is extremely high. The solution then seems obvious: acquire more information. However, even if the value of information is high *once it is acquired*, the expected value of *trying to acquire* more information is much lower, because it is actually hard to acquire that information.

Given our current epistemic state, I cannot say which option is the better one: reducing extinction risk or trying to acquire more information. However, I *do* say that we should not let these considerations paralyze us; either one or the other option is better. It is tempting to sit on our hands in the face of large uncertainty, because it *seems* to exempt us from blame. Instead, we should act on our best guess, EMV, how fragile it may be. If longtermism is correct, our influence on the long-term future is incredibly important. We should use any evidence we can get to inform our decisions not despite, but because of the stakes at hand.

Luckily, we might not be in such a dire situation as described here. In real life, there is no extinction button and decisions are more complex. Interestingly, adding complexity could make our decisions *easier* rather than more complicated. In the next chapter, I discuss a number of approaches to the practical problem of whether we should extinction risk. After first ruling out some suggested approaches, I make the case that reducing extinction risks possibly has significantly large and positive side effects on the value of the long-term future.

Chapter 4. Given our uncertainties, is it good to reduce human extinction risk?

The previous chapter looked at problems with assessing the expected moral value of the entire future. As noted in chapter 1, this was a simplification to highlight some philosophical problems. If we want to know what we ought to *do*, we should look instead at how the actions available to us would influence the expected moral value of the long-term future. In this chapter, I discuss our possible actions from two perspectives. In section 4.1, I discuss 'pragmatic methods' that are commonly used to deal with uncertainty in daily life. After arguing that these methods do not offer a solution, I turn to the expected effects of reducing extinction risk and discuss whether side-effects of reducing extinction risk should make us more positive about the expected moral value of reducing human extinction risk.

4.1 Pragmatic methods to deal with moral uncertainty

Given that we are uncertain about the expected value of the future and about what we ought to do, it is tempting and sensible to search for pragmatic ways to deal with uncertainty. In daily life, we have at least the following actions available:

> *Preserve Option Value.* Invest in options that leave open many possibilities, which is beneficial when these possibilities turn out to be valuable. In contrast, avoid irreversible decisions, which may accidentally exclude highly valuable paths permanently.

> *Acquire Information.* Reduce or resolve uncertainty by *acquiring more information.*

And if those are unavailable, we might opt for the following:

> *Do Nothing.* Go with the default option and do not intervene in how events unfold.

If any of these options are available and justified, then we would not have to deal (yet) with the difficult philosophical problems from the previous chapters. However, I argue that none of these traditional options are possible or satisfying in this context.

4.1.1 Option value: neither clearly positive, nor big

Some people have suggested we should reduce existential risk for its *option value* (Bostrom, 2013; MacAskill, 2014). Bostrom (p. 24) writes:

> If we are indeed profoundly uncertain about our ultimate aims, then we should recognize that there is a great *option value* in preserving - and ideally improving - our ability to recognize value and to steer the future accordingly. Ensuring that there will be a future version of humanity with great powers and a propensity to use them wisely is plausibly the best way available to us to increase the probability that the future will contain a lot of value. To do this, we must prevent any existential catastrophe.

Remember that an existential catastrophe is "the extinction of Earth-originating intelligent life or the permanent and drastic failure of that life to realise its potential for desirable development" (Bostrom, 2013, p. 15). Since Bostrom includes all extinction events as existential catastrophes, I am focusing my criticism on the argument that reducing *extinction risk* has great option value. To criticize the argument, let me first deconstruct the argument into four premises and a conclusion.

Premise 1: We are profoundly uncertain about our ultimate aims.

Premise 2: If we are profoundly uncertain about our ultimate aims, then we should recognize that there is a great option value in preserving - and ideally improving - our ability to recognize value and to steer the future accordingly.

From *Premise 1* and *2* follows:

Premise 3: There is great option value in preserving - and ideally improving - our ability to recognize value and to steer the future accordingly.

Premise 4: Preventing [extinction] preserves - and ideally improves - our ability to recognize value and to steer the future accordingly.

Conclusion: Preventing [extinction] has great option value.

Preventing extinction (and other existential catastrophes) probably ensures that there "will be a future version of humanity with great powers" (assuming technological development will continue). However, although preventing extinction is necessary to ensure that our descendants will have "a propensity to use [their great powers] wisely", it is not sufficient. We cannot assume that our descendants will necessarily be wise and altruistic without argument.[33] As a consequence, preventing extinction also leaves the option open that the future will contain a lot of negative value, because great power might be combined with a lack of wisdom or coordination. In what follows, I will criticize *Premise 2*: that when we are uncertain, there is great option value in preserving our ability to recognize value and steer the future accordingly.

How would 'preserving our ability to recognize and steer the future' yield option value? Normally, the option value of an asset is high when there is large uncertainty about the future need of the asset, and when losing the asset is irreversible (or comes with high costs). In this case, human civilization is the asset. Both conditions seem to be met; there is uncertainty about whether human civilization will be a positive influence on the value of the future, and extinction is mostly irreversible.[34] However, a third factor affecting option value is the extent to which one has the future ability to choose an option based on more information. This is where the argument is weakest.

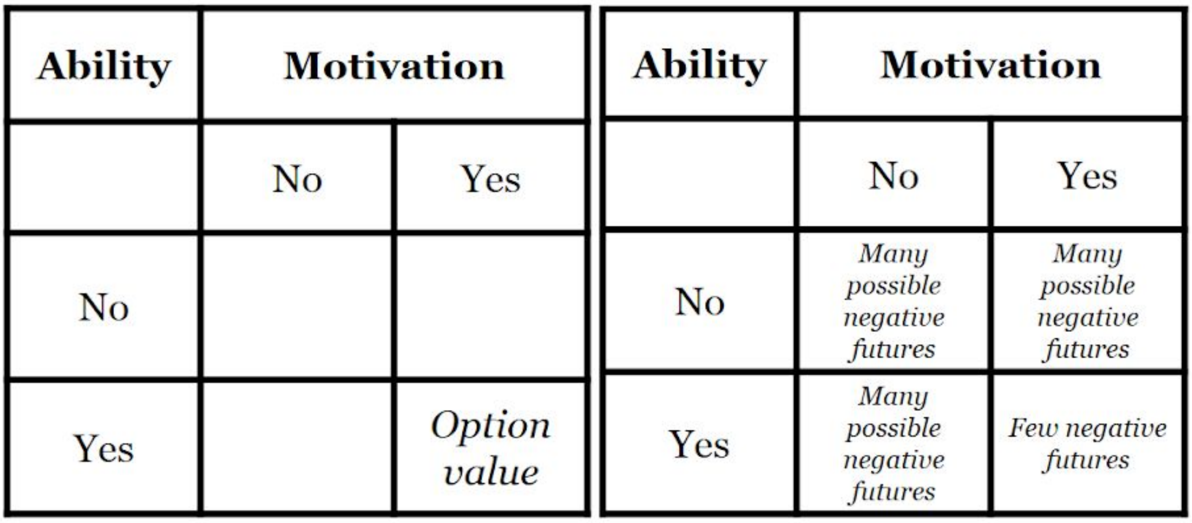
Suppose we postpone extinction. Can future generations choose to change the course of the future if information is available that the expected value of the future is negative? Would humanity go as far as choosing extinction if the future looks bleak, as MacAskill (2014, p. 240) suggests humanity can?[35] Let's survey the possibilities. A future version of humanity is either capable or incapable to significantly change trajectory if it wants to[36], and either motivated or unmotivated to change trajectory of the expected moral value of the future looks negative. Below, in the left table, we see where option value resides: when humanity is motivated and able to change trajectory. In the right table we see where to expect the future to be negative.

[33] In the above quote, Bostrom (2013) does not literally state that preventing existential catastrophe *ensures* that there will be a future version of humanity with great powers and a propensity to use them wisely, only that preventing existential catastrophe is *necessary*. However, he does not address the possibility of preventing existential catastrophe resulting resulting in an unwise future version of humanity anywhere in the paper.

[34] Given that Earth will remain hospitable to complex life for approximately a few hundred million to a billion years (O'Malley-James, J. T., Cockell, C. S., Greaves, J. S. and Raven, 2014), it is possible that another intelligent and complex civilization arises in that timespan. Thus, the capabilities lost by extinction of humanity are not irreversible for certain. On the other hand, extinction is not reversible in the sense that one can make a choice to reverse the situation based on new information.

[35] MacAskill (2014, p. 240) writes "If we continue to exist, then we always have the option of letting ourselves go extinct in the future (or, perhaps more realistically, of considerably reducing population size)."

[36] Ability is probably a combination of ability to alter the physical environment + ability to coordinate with other agents.



Ability	Motivation	
	No	Yes
No		
Yes		*Option value*

Ability	Motivation	
	No	Yes
No	*Many possible negative futures*	*Many possible negative futures*
Yes	*Many possible negative futures*	*Few negative futures*

Figure 6a (left) and 6b (right). Possible combinations for a future version of humanity. ‘Ability’ stands for ‘ability to significantly change the course of the future if they want to’. ‘Motivation’ stands for ‘will want to significantly change the course of the future if it looks to have negative expected moral value’. Most of the option value resides in the scenarios in which the future looks very positive.

Only when humanity is both able and motivated to significantly change the course of the future do we have option value. However, suppose that our descendants both have the ability and the motivation to affect the future for the good of everyone, such that a future version of humanity is wise enough to recognize when the expected value of the future is negative and coordinated and powerful enough to go extinct or make other significant changes. As other authors have raised (Brauner & Grosse-Holz, 2018), given such a state of affairs it seems unlikely that the future would be bad! After all, humanity would be wise, powerful, and coordinated. Most of the bad futures we are worried about do not follow from such a version of humanity, but from a version that is powerful but unwise and/or uncoordinated.

To be clear, there would be a small amount of option value. There could be some fringe cases in which a wise and powerful future version of humanity would have good reason to expect the future to be better if they went extinct, and be able to do so. Or perhaps it would be possible for a small group of dedicated, altruistic agents to bring humanity to extinction, without risking even worse outcomes. At the same time they would need to be unable to improve humanity’s trajectory significantly in any other way for extinction to be their highest priority. Furthermore, leaving open this option also works the other way around: a small group of ambitious individuals could make humanity go extinct if the future looks overwhelmingly positive.

In conclusion, deferring our choice to continue or not to our descendants yields little option value. In most of the scenarios in which they could decide to altruistically go extinct (or otherwise change the course of the future) it will not be needed, precisely because they would be altruistic and capable enough that the future would look bright and promising.

4.1.2 We cannot wait to resolve moral uncertainty and disagreement

Another way to deal with uncertainty is simply to reduce it by acquiring more information and knowledge. More information has value when it can change what we ought to do. However, acquiring useful information is difficult in this case. There are two large sources of uncertainty that influence the expected value of the long-term future. First, there is empirical uncertainty regarding what the future will be like. This is a matter about which few confident conclusions can be drawn. Therefore we will, for a long time, remain empirically uncertain about the long-term future. The second source is moral uncertainty: humanity, and philosophers in specific, do not agree on what is valuable or what we ought to do. And the field that deals with different sizes of populations, *variable population ethics*, is especially rich with impossibility theorems: every known population axiology has implications that a significant subset of ethicists find unacceptable and repugnant (Greaves, 2017). We should not expect resolution of this disagreement anytime soon.

4.1.3 Doing nothing is also a choice

In light of all this uncertainty, one is tempted to simply throw their hands up and go with the default option: doing nothing and ignoring the issue. However, as is often the case, doing nothing is also a choice. In fact, it is favoured by some theories, and disfavoured by others. Consider a theory which implies that going extinct is good because it implies that the future is likely to be bad. Such a theory prefers inaction when existential risk is significant, because eventually extinction will occur. However, such inaction is horrific according to a theory which implies that going extinct is an enormous tragedy, because it implies that the future is likely to be enormously valuable. Therefore, ‘doing nothing’ should not receive the special status of a default option that we can opt for if we do not manage to deal with moral uncertainty. Instead, it should enter our decision procedure as just another option we have, and therefore it is not a pragmatic solution to dealing with moral uncertainty.

Neither of the above three options succeeds as a pragmatic solution to dealing with moral uncertainty. Instead, we should engage with our moral uncertainty we have about the possible actions we can take to influence the long-term future, which I do in the next section.

4.2 Does reducing extinction risk have positive expected value?

4.2.1 Focus on the EMV of trying to reduce human extinction risk

So far, we have looked at the expected moral value of the entire long-term future from the perspective of a non-influential observer and tried to compare it to the expected moral value of a possible α-future in which no moral theory sees any value. We did this to highlight some philosophical issues with assessing the EMV of the entire long-term future.

However, it would be mistaken to conclude anything about the EMV of actions intended to reduce human extinction risk solely on the basis of the EMV of the entire long-term future. Instead, we should compare the EMV of the long-term future conditional on 'no intervention' to the EMV of the long-term future conditional on additional extinction risk reduction. To do this, we do not have to assess the EMV of either scenario in its totality; we only have to estimate the difference between the two values. In other words, we have to estimate whether additional effort to reduce human extinction risk, in expectation, increases the EMV of the long-term future.

4.2.2 A primer on extinction risks

At this point it would be good to step away from the abstract and discuss some realistic ways that could lead to extinction of humanity in the next few centuries. This helps us to understand why it is good (in expectation) to address these risks, even given uncertainty about the expected moral value about the long-term future.

The most familiar sources of human extinction risk are probably climate change, nuclear war, and large asteroid impacts. For our purposes it is easiest to structure these and other risks according to their possible path to extinction. Let's first discuss indirect paths to extinction: paths in which humanity first (quickly or slowly) collapses, and then goes extinct. For example, both nuclear war, large asteroid impacts, and supervolcanic eruptions probably do not lead to direct extinction (Tonn & MacGregor, 2009), but first lead to civilizational collapse via their consequences: an *atmospheric winter* (Baum, Denkenberger, Pearce, Robock, & Winkler, 2015). This results from many dust and ash particles suspended in the atmosphere for years, blocking sunlight. As a result, temperature drops significantly with regular frosts throughout the year in most areas. As a result, it is hard to grow food and most people would die from starvation. Another path is a more general *system collapse*. This occurs when a critical system is destroyed and results in many other systems unable to function, and consequently also collapse (Pamlin & Armstrong, 2016). For example, increasing dependence on electricity makes the global system vulnerable if the electrical grid is destroyed (e.g. by solar winds). We have come to depend on electricity for important communication and transportation. If that is disrupted, it will consequently disrupt economic systems such as food production and distribution, resulting in mass starvation. Other examples that can lead to collapse are climate change, great power war, and extreme financial crisis (Pamlin & Armstrong, 2016). A third risk of collapse comes from *extremely deathly scenarios*, most prominently a virulent and deathly pandemic or weapons of mass destruction. These could directly kill most, if not all, humans instead of destroying systems humans rely on.

Although collapse of human civilization is horrible, it is unclear whether collapses are likely to lead to extinction. Torres (2017) mentions the *last few people question*:

> Given that a (very) large proportion of humanity dies off, how likely is it that the last few survivors die out, rather than form the start of recovery towards a new technologically advanced civilization?

Upon reflection, most scholars and futurists (Baum et al., 2019; Bostrom, 2013; Maher & Baum, 2013) have concluded that it is not unlikely that a relatively small group of humans can succeed in (partially) rebuilding civilization, given enough time. For example, humans relying on fishing, hoarded food stockpiles, or edible mushrooms can survive the initial transition period well-enough to maintain a viable and genetically diverse 'founding population.' From such a founding population, it seems reasonable to assume technology will develop rather than stagnate, given enough time. In conclusion, although modern civilization is vulnerable, humanity as a whole seems rather resilient to collapse. However, this does not mean that recovery is inevitable, nor that recovery is expected to lead to equally favorable conditions as the current global civilization (Beckstead, 2015). This becomes an important point in the next section.

Besides collapse, there is the possibility of direct extinction, or very rapid collapse that rapidly leads to extinction (the distinction is blurry). Because humanity appears quite robust, we need to consider emerging technologies for scenarios of direct extinction. Advancing biotechnology and advanced artificial intelligence (Bostrom, 2014b) may be the largest factors of extinction risk.

Future advances in biotechnology could lead to cheap and accessible technology to create ‘super viruses.’ If these viruses are released, either accidentally or purposefully (Millett & Snyder-Beattie, 2017b), a significantly potent virus could spread wide enough to bring humanity beneath a minimum viable population size (Millett & Snyder-Beattie, 2017a) and lead to extinction.

Another ‘direct’ extinction risk is superintelligent AI. Intelligence is the reason for humanity’s dominance on Earth; outsourcing important decisions to smarter-than-human AI should therefore be done carefully, lest it shapes the world in undesirable ways (Bostrom & Yudkowsky, 2018). However, AI risk is different from most extinction risks in two ways: first, whereas most extinction risk sources are more likely to lead to collapse than extinction, this is less clear for AI. A misaligned superintelligent AI could pursue the instrumental goal of wiping out humans to maximize the probability that it achieves its final goal(s) (Bostrom, 2014b; Omohundro, 2008). Such an intentional pursuit is much more likely to lead to extinction than the ‘dumb’ processes described earlier, because the ‘last few survivors’ can be actively sought out. A superintelligent AI could transform its environment so dramatically that humans can no longer live in it. A second difference between AI risk and other sources of extinction risk is that AI can lead to both much worse outcomes than extinction if misaligned with our values (cf. chapter 1) or much better outcomes, as well as reduce extinction risk to very low levels if it is aligned with our values. I believe this makes the alignment of advanced AI important to address for practically every ethical theory (cf. Sotala & Gloor (2017) for an argument that Asymmetry-like moral views should prioritize AI alignment). However, it is beyond the scope of this thesis to survey the arguments for whether we should take risk from AI seriously. In the next section, I will focus on factors other than AI that could lead to extinction.

4.2.3 Positive side-effects of reducing non-AI sources of extinction risk

Many risk sources that could lead to extinction could also lead to other outcomes. To assess the value of reducing extinction risk, we should assess the possible consequences of a risk actualizing (e.g. a nuclear exchange happening), estimate the probabilities of each possible consequence, and assess the EMV of each possible consequence compared to the risk source not actualizing. Let’s start with the possible consequences. For most extinction risk sources, we can classify the possible consequences as follows:

Extinction	Direct extinction or collapse followed by extinction
Recovery	Civilizational collapse followed by recovery to our current technological level
Global Disruption	Global systems are disrupted, but there occurs no civilizational collapse

To assess the expected moral value of extinction risk reduction, we need to assess the relative probabilities of these three types of consequence. First off, let's look how the probabilities *rank*. We can stipulate that global disruption is more likely than both extinction and recovery; when events are caused by a similar underlying system, small events are more likely than bigger events (Clauset, Shalizi & Newman, 2009). For example, there are many more smaller storms than massive hurricanes, and they are generated by a similar underlying system: the weather system. Similarly, we can stipulate that collapse is more likely than direct extinction[37] because it is a 'smaller' event in terms of impact. More contentious is whether recovery is more likely than extinction. Based on the last few people question from the previous section, I assume it is.

Going beyond the ranking, it is more difficult to assess whether there are large or small difference in the relative probabilities of the different types of consequences. On the one hand, we might expect there to be orders of magnitude differences: global disruption would be ten or a hundred times more likely than recovery, and recovery ten or a hundred times more likely than extinction. However, if it turned out that civilization is fragile, such that a risky event leads either to no change (because they are too small) or to civilizational collapse, then the relative probabilities of global disruptions and recovery are much closer together. In addition, if recovery from civilizational collapse is actually not very likely (e.g. because regaining industry requires fossil fuels which are no longer available for a recovering civilization[38]), there is only a small difference between the probability of extinction and recovery.

[37] It is a difficult question when to classify a series of events as *direct extinction* versus *indirect extinction*. For example, when the last surviving population after a global nuclear exchange dies twenty years after the exchange from starvation, is that direct extinction because the exchange was so severe it reduced the remaining populations to critically small groups, or is it a failure of recovery? However, we assume that such a distinction can be made, because not much turns on exactly *where* we draw the line.

[38] This point was raised by Karim Jebari in a presentation on the likelihood of civilizational recovery after collapse at the Effective Altruism Global (X) Nordics conference in 2019.

Thus, roughly speaking, one of the following two models is correct, but we do not know which one:

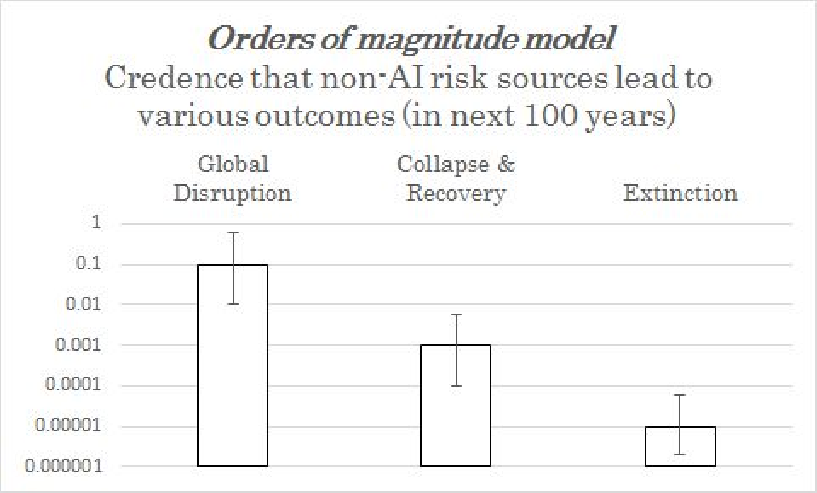


Figure 7a. Visualisation of the assumption that sources of extinction risk are more likely to lead to less severe outcomes than more severe outcomes.

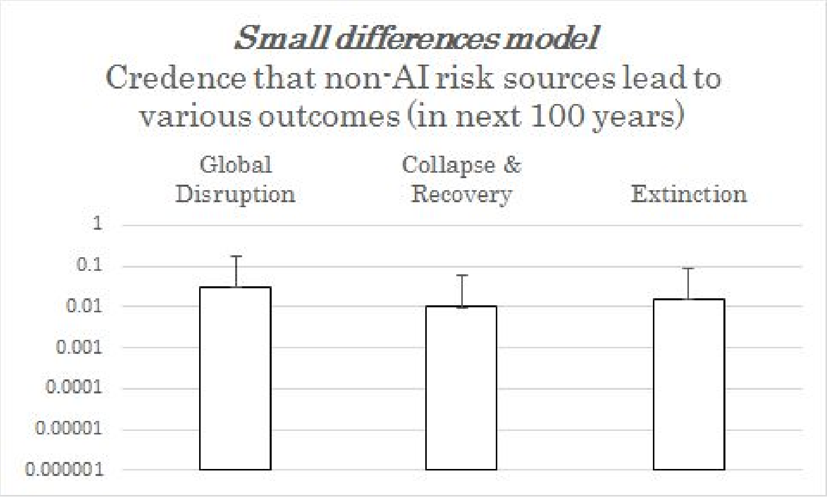


Figure 7b. Visualisation of the assumption that sources of extinction risk are similarly likely to lead to less severe outcomes than more severe outcomes.

Which model is correct significantly influences whether or not we should human reduce extinction risk. Before we continue to the implications of each model, we first need to assess the expected moral value of each type of consequence. To keep things simple, we assume the earlier 50/50-credence distribution over Totalism and a Moderate Asymmetric

View with an N-ratio of $\frac{1}{10}$.[39] To start with extinction, we assume that it is roughly as good or bad as a universe devoid of moral value.

Whether recovery after collapse (if it happens) is in expectation better or worse than our current trajectory appears more speculative. We could look for *'critical junctures'* in our trajectory. If we can identify junctures where we have obviously taken the path that is worse for our long-term future, this should make us more optimistic about a retry. If we can identify critical junctures where we have obviously taken the right path but we might have chosen differently, this should make us wary about a retry. It is far beyond the scope of this thesis to assess these matters, but I find it personally very hard to identify particularly important junctures where humanity has obviously taken a right or wrong turn. Furthermore, one needs to take into account how different starting conditions affect a recovery trajectory. For example, how would a second trajectory be affected by fewer natural resources like oil, by higher carbon dioxide levels, or by leftover knowledge from our current civilization? This discussion is also shaped by whether one thinks technological and cultural evolution have a 'direction': if one thinks it does not (e.g. along the lines of Gould's (1994) arguments against biological evolution having a direction), collapse seems unlikely to lead to recovery and the desirability of recovery will be highly unpredictable. If one thinks it does, for example a tendency towards more complexity (cf. Tainter, 1995 for an argument in support of this view) or more cooperation (cf. Wright (2001) for an argument in support of this view), then it becomes easier to predict the desirability of recovery, given some changed starting conditions.[40] In conclusion, given our ignorance, we should not expect recovery to be significantly better or worse than our current trajectory.

However, more can be said about whether global disruption is good or bad (in expectation) for the value of the future.[41] This is speculative, but Beckstead (2015) lays out some reasons to expect that global disruption will put humanity most certainly on a worse trajectory: it may reverse social progress, limit the ability to adequately regulate the development of dangerous technologies, open an opportunity for authoritarian regimes to take hold, or increase inter-state conflict. All of these increase the probability of suffering risks. We can also approach the issue abstractly: disruption can be seen as injecting more noise into a previously more stable global system, increasing the probability that the world

[39] Since these arguments are more qualitative in nature and do not go into specifics, exact credences and theories are not actually necessary. The considerations would be the same if more (unbounded) theories would be included.

[40] Lukas Gloor (2018) believes extinction risk reduction is negative, or at least not the most promising option for downside-focused ethical theories. However, he does not note global disruption, and only discusses the desirability of recovery after collapse.

[41] Note that *specific* global disruption (e.g. disruption of science, or disruption of wealth distribution) has more predictable effects than *general* global disruption, but it seems more reasonable to assume the disruption is global rather than specific, and even if it is specific, it is difficult to predict which system will be disrupted.

settles into a different semi-stable configuration. If there are many more ways for the world to become worse than better, increasing randomness is more likely to lead to an undesirable state of the world. I am convinced that, unless we are currently in a particularly bad state of the world, global disruption has (in expectation) a very negative effect on the value of the long-term future for a wide range of moral views.

If these values are correct, then reducing extinction risk would have positive expected value if global disruption is orders of magnitude more likely than the other scenarios. However, if global disruption is not actually that much more likely than extinction, Asymmetric Views might judge reducing extinction risk as negative expected value compared to letting things develop without intervention. It lies beyond the scope of this thesis to assess whether the differences in relative probabilities are small or large. A more quantitative model - also beyond the scope of this thesis - is better suited to assess under which conditions the above considerations turn out in favour of extinction risk reduction.

Chapter 5. Conclusion

5.1 Summary & conclusion

Influencing the risk of human extinction is plausibly the most effective lever to influencing the long-term future. I focused on two central questions: "Can we expect the future to be good?" and "Is reducing extinction risk good?" To answer the first question, I first looked at the philosophical issues in assessing the expected moral value of the long-term future. I showed that moral uncertainty between unbounded moral theories forces us to engage with the problem of intertheoretic comparison. I considered further objections to applying expected moral value to cases with large population but rejected them. Then, I discussed why some pragmatic options fail to deal with our current moral uncertainty; I rejected the argument from option value, dismissed the hope that we could resolve our uncertainty, and dismissed inaction as a safe default option. Afterwards, to answer the second question, I identified two models for the value of extinction risk reduction. In the first model, the probabilities of different outcomes (extinction, recovery after collapse, and global disruption) differ by orders of magnitude. In the second model, the probabilities of different outcomes differ only by a little.

What do these results mean? Can we say whether, given our uncertainty, the value of the future is positive or negative and is it clear whether we should reduce extinction risk? Unfortunately, I cannot give a straight answer; it depends. If one has a high credence in Totalism, if one is comfortable with fragile conclusions, and if one believes intertheoretic value comparisons are fundamentally possible, then the future looks positive. Otherwise, all bets are off. Whether extinction risk reduction has positive expected value depends in large part on how much more likely it is that various extinction risk sources would lead to global disruption rather than extinction. Since global disruption seems clearly negative from practically all moral views, reducing the risk from these sources would be positive if global disruption was orders of magnitude more likely than extinction. However, I cannot assert with certainty that the orders of magnitude model is correct. This should only be a worry for those who are on the fence or pessimistic about the value of the future; if one confidently believes that extinction is worse than non-extinction, extinction risk reduction is clearly positive.

5.2 Theoretical implications

In contrast to earlier work by Greaves and Ord (2017), I have shown that in large population cases the expected moral value approach does not necessarily defer to the preferences of Totalism. Specifically, Asymmetric Views that weigh the creation of net-negative lives more heavily in their value function than the creation of net-positive lives sometimes overwhelm Totalism instead. In practice, which theory dominates depends on the particular normative and expected empirical ratios, as well as how both theories are normalized to make them comparable. Although this is in line with Greaves' and Ord's more general point that EMV favours unbounded theories in large population cases, it also shows moral uncertainty still plays a role in these cases, even if it is a smaller role because only uncertainty between unbounded views need to be considered. However, some scholars may take this property of the EMV approach (i.e. that unbounded views dominate in large population cases) as a reason to reject or modify the EMV approach and further work could be done to propose viable alternatives. Moreover, including Asymmetric Views dampens optimism about the long-term future: it may be very negative according to some moral views.

I have also brought up the problem of *fragile EMV*. I take the position that this does not fundamentally change how we should make decisions; it just increases the value of information. However, other scholars may disagree and may wish to explore this issue further in the context of problems for orthodox Bayesianism.

5.3 Practical implications

If extinction risk reduction is valuable, it is a good candidate to be humanity's current top priority. For individuals, this means they could donate to extinction risk reducing institutions, tailor their career to extinction risk reduction, or exert pressure on their democratic representatives to reduce extinction risk. Governments could set up representation for future generations and cooperate internationally to address global catastrophic and existential risks (cf. Farquhar, Cotton-Barratt, Halstead, & Schubert (2016) for concrete proposals).

However, there are alternative strategies of improving the long-term future that may be higher priority, such as moral circle expansion (i.e. including more morally relevant entities in our 'circle of concern', Reese, 2018; Singer, 1981), promoting peace, improving institutional decision-making, or reducing the risk of totalitarianism (Caplan, 2008). Which strategy is most promising depends on the specifics; besides being tractable, any feasible candidate strategy needs to have both very long-run or persistent effects, and be urgent to address currently rather than later. For example, if we expect continued moral progress, promoting moral circle expansion will only speed up this process, instead of having long-lasting effects. There are a few other practical implications of moral uncertainty about the long term-future.

Groups with different value systems should cooperate

Most of this thesis has focused on moral *uncertainty within* an agent, rather than moral *disagreement between* different agents. Although there is significant overlap in these cases, there are also substantial differences. A single rational agent will not sabotage themselves because of uncertainty. However, multiple disagreeing rational agents can all act rationally, yet achieve suboptimal outcomes by drifting into a Nash equilibrium.[42] For example, if one agent wants to increase the probability of a controversial trajectory, while another wants to decrease it, they both waste their resources. Instead, they could engage in *moral trade* (Ord, 2015): both committing to leave the controversial trajectory alone and spend their resources in more mutually beneficial ways.

Reduce empirical uncertainty

For many people, the expected moral value of reducing extinction risk is highly uncertain. Therefore, more information about the possible and probable values of the relevant parameters is extremely important. However, more information is hard to come by. If we strive to reduce our uncertainty, should we prioritize empirical or moral sources of uncertainty? Reducing either kind of uncertainty has complementary effects; if we know what is valuable we know which empirical information is important. Vice versa, if we know with more certainty how the future is likely to go, we know which philosophical theories will actually disagree on what we ought to do. However, I argue that reducing empirical uncertainty is a better way to reduce our uncertainty.

First, moral philosophy has so far not converged on a particular moral theory and has instead diverged, branching out into more and more theories to be considered. Some may see that as the result (or fault) of moral methodology, while others see it as evidence against an underlying moral truth or a single correct moral theory. In either case it is evidence that the current methodology in moral philosophy is unlikely to significantly reduce our moral uncertainty. On the other hand, I do not imply that moral philosophy has not made progress. However, this progress seems to come more often from considering

[42] An oft-quoted example is the Prisoner's Dilemma, in which two people are interrogated for a crime they committed. Even though both people would be better off if they both kept silent (i.e. Cooperate), they are both incentivized to tell on the other (Defect) no matter what the other chooses. As a result, two rational agents would end up in Defect-Defect in a single-shot Prisoner's Dilemma: a highly suboptimal outcome.

areas or problems that previously have not been considered or by incorporating new empirical information. It seems much rarer that, once different sides have built their theories and dug in, doing moral philosophy significantly reduces moral uncertainty, even though it may clarify it.

Second, I believe there have been few serious attempts to estimate the expected moral value of reducing extinction risk that take into account the probability of very negative futures. There have been some impressive individual works exploring possible futures (cf. Bostrom, 2014b; Hanson, 2016; Tegmark, 2017), but there has not been a large and dedicated research project combining all considerations.[43] This is either evidence that there are some easy pickings left, or that many individuals have concluded the project to be hopelessly intractable. However, I believe that a serious and significant effort is the only way to find out whether it is impossible, or merely very difficult to more precisely estimate the expected moral value of the long-term future. Therefore, I believe there is good reason to try such a research project, although this is not sufficient to show that it should be: it might not be a higher priority than reducing extinction risk, and a bad attempt could be worse than no attempt at all.

5.4 Limitations and further research

There are a few limitations to this thesis. First, I only considered some consequentialist axiologies: Totalism and Personal Moderate Asymmetric Views. Of course, there are many more views that have something to say about the long-term future. Some of them will often disagree with Totalism (e.g. Strict Asymmetric Views), while others cannot be incorporated into the EMV procedure because they resist being represented by cardinal rankings. For example, rights-based theories do not assign values to options, which makes them hard to compare to consequentialist theories under moral uncertainty. Further research could explore the effect of including these other moral theories on what we ought to do for the long-term future.

Another limitation is that I made the implicit assumption that the risk of humanity's extinction is not miniscule, but rather significant enough that it is worthwhile to address. If it was very small, further reductions would be difficult and costly to make, and actions that could improve the *quality* of life in the future would become higher priority (e.g. moral circle expansion, improving institutional decision making).

[43] The largest research projects I am aware of are the totalism-leaning *Future of Humanity Institute* (University of Oxford) and the smaller suffering-focused *Foundational Research Institute*. However, both institute are young, small, and are, at the time of writing, prioritizing how to affect the future - primarily by reducing existential risks and suffering risks respectively.

Furthermore, Personal Asymmetric Views, the main rival for Totalism according to this thesis, rely on a well-defined concept of 'lives.' However, as technology advances the line may blur between 'creating a new life' and 'enhancing current life.' For example, if digital sentience is possible, would creating a digital clone be 'creating new life'? How about situations in which part of our 'neural hardware' is shared with other beings? Such technological advances may show that our concepts of life and identity are too vague to be morally relevant or useful.

As a last limitation, I did not create quantitative models for estimating the value of reducing extinction risk. It is possible that for some parameter values (e.g. a low probability of recovery after collapse combined with a negative expectation of the future) extinction risk reduction is negative, even taking into account the positive side effects.

There are a number of empirical research topics this thesis touched on that would be valuable to research further. First, there is much more research needed to address the sources of extinction risk, and this research is plausibly the highest priority. However, there are also other research avenues that would reduce our uncertainty about the EMV of the long-term future. For example, it is still highly unclear how recovery after civilization's collapse might turn out and how likely recovery after collapse is. Research addressing this would need to draw from - among others - history, economics, sociology, philosophy, and human geography (cf. Maher & Baum (2013), Baum & Tonn (2015) and Jebari (Unpublished) for earlier work on this topic).

Furthermore, it is not clear how large the differences in likelihood are of global disruption, collapse, and extinction. If they are small, the positive side effects might not weigh up against the positive expected value of extinction if one believes the future is significantly negative. Further research into the disruptive effect of global catastrophes would be highly valuable. More research would help also to provide stronger arguments to persuade large actors to address these issues.

Lastly, although there have been some estimates about how large the future can be (Bostrom, 2003), it is still unclear how the probability mass should rationally be distributed over various sizes of the future. This is important because large futures tend to dominate the expected moral value unless they are proportionally improbable. That is, unless each order of magnitude in size is an order of magnitude less probable.

All of these approaches would contribute to, possibly, the most important project there is: creating a valuable long-term future for humanity and all sentient life.

References

Adams, F. C. (2008). Long-term astrophysical processes. In N. Bostrom & M. M. Cirkovic (Eds.), *Global Catastrophic Risks* (pp. 33–47). Oxford University Press.

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

Althaus, D., & Gloor, L. (2016). Reducing Risks of Astronomical Suffering: A Neglected Priority. *Foundational Research Institute: Berlin, Germany.*

Arrhenius, G., Bykvist, K., Anderberg, T., Carlson, E., Pol, T., Ryman, P., … Österberg, J. (1994). Future Generations and Interpersonal Compensations Moral Aspects of Energy Use. *Uppsala Prints and Preprints in Philosophy,* (21).

Baron, J., & Greene, J. (1996). Determinants of Insensitivity to Quantity in Valuation of Public Goods: Contribution, Warm Glow, Budget Constraints, Availability, and Prominence, *2*(2), 107–125.

Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., … Yampolskiy, R. V. (2019). Long-Term Trajectories of Human Civilization. *Foresight,* 1–34.

Baum, S. D., Denkenberger, D. C., Pearce, J. M., Robock, A., & Winkler, R. (2015). Resilience to global food supply catastrophes. *Environment Systems and Decisions,* *35*(2), 301–313.

Baum, S.D., & B.E. Tonn. (2015). Confronting Future Catastrophic Threats to Humanity. *Futures,* 72 (September): 1–3.

Beckstead, N. (2013). On the overwhelming importance of shaping the far future. *ProQuest Dissertations and Theses,* 199.

Beckstead, N. (2015). The long-term significance of reducing global catastrophic risks. *The GiveWell Blog.* Retrieved from https://blog.givewell.org/2015/08/13/the-long-term-significance-of-reducing-global-catastrophic-risks/

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

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

Bostrom, N. (2003). Astronomical Waste : The Opportunity Cost of Delayed Technological Development. *Utilitas, 15*(3), 308–314.

Bostrom, N. (2009). Pascal's mugging. *Analysis, 69*(3), 443–445.

Bostrom, N. (2009). The Future of Humanity. *New Waves in Philosophy of Technology,* 551–557.

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

Bostrom, N. (2014a). Crucial Considerations and Wise Philanthropy. Retrieved from https://www.effectivealtruism.org/articles/crucial-considerations-and-wise-philanthropy-nick-bostrom/

Bostrom, N. (2014b). *Superintelligence: paths, dangers, strategies.* Dunod.

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence,* 316–334.

Brauner, J. M., & Grosse-Holz, F. (2018). The expected value of extinction risk reduction is positive. *Effective Altruism Forum.* Retrieved from https://www.effectivealtruism.org/articles/the-expected-value-of-extinction-risk-reduction-is-positive/

Briggs, R. A. (2017). Normative Theories of Rational Choice: Expected Utility. Retrieved from https://plato.stanford.edu/archives/spr2017/entries/rationality-normative-utility/

Bykvist, K. (2017). Moral uncertainty. *Philosophy Compass, 12*(3), 1–8.

Caplan, B. (2008). The totalitarian threat. In N. Bostrom & M. M. Cirkovic (Eds.), *Global Catastrophic Risks* (p. 498). Oxford University Press.

Caviola, L., & Schubert, S. (Unpublished raw data). Psychology of Existential Risk and Longtermism.

Clauset, A., Shalizi, C. R. & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review.* 51 (4): 661–703.

Cotton-Barratt, O., & Ord, T. (2015). Existential Risk and Existential Hope: Definitions. *Future of Humanity Institute, 1,* 1–4.

Farquhar, S., Cotton-Barratt, O., Halstead, J., & Schubert, S. (2016). Existential Risk: Diplomacy and Governance, 1–34.

Gaba, J. M. (1999). Environmental ethics and our moral relationship to future generations: Future rights and present virtue. , 24, 249. *Colum. J. Envtl. L., 24,* 249–288.

Gardiner, S. M. (2009). A contract on future generations? *Intergenerational Justice,* 77–118.

Gloor, L. (2018). Cause prioritization for downside-focused value systems. *Foundational Research Institute: Berlin, Germany.*

Gould S.J. (1994): The Evolution of Life on Earth, *Scientific American, 271 (4), p. 62-69.*

Greaves, H. (2016). Cluelessness. *Proceedings of the Aristotelian Society, 116*(3), 311–339.

Greaves, H. (2017). Population ethics and population axiology: The basic questions. *Philosophy Compass, 12*(11).

Greaves, H., & Ord, T. (2017). Moral uncertainty about population ethics. *Journal of Ethics and Social Philosophy, 12*(November), 135–167.

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

Hedden, B. (2016). Does MITE Make Right? On Decision-Making under Normative Uncertainty in Shafer-Landau (Ed.) *Oxford Studies in Metaethics 11.*

Jebari, K. (Unpublished). Resetting the tape of history: what can we infer about history from instances of convergent cultural evolution?

Joyce, J. M. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, (19).

Knutsson, S. (2016). Measuring Happiness and Suffering, (May), 70–71. *Foundational Research Institute: Berlin, Germany.* Retrieved from https://foundational-research.org/files/measuring-happiness-and-suffering.pdf

Leslie, J. (2002). *The End of the World: the science and ethics of human extinction.* Routledge.

MacAskill, W. (2014). *Normative Uncertainty.* University of Oxford.

Maher, T. M., & Baum, S. D. (2013). Adaptation to and recovery from global catastrophe. *Sustainability (Switzerland), 5*(4), 461–1479.

Mayerfeld, J. (1996). The Moral Asymmetry of Happiness and Suffering, *The Southern journal of philosophy, 34*(3), 317-338.

McMahan, J. (2009). Asymmetries in the morality of causing people to exist. In M. A. Roberts & D. T. Wasserman (Eds.), *Harming future persons* (pp. 49–68). Springer.

Meyer, L. H. (2017). *Intergenerational Justice.* Routledge.

Millett, P., & Snyder-Beattie, A. (2017a). Existential Risk and Cost-Effective Biosecurity. *Health Security, 15*(4), 373–383.

Millett, P., & Snyder-Beattie, A. (2017b). Human Agency and Global Catastrophic Biorisks. *Health Security, 15*(4), 335–336.

Morgenstern, O., & Von Neumann, J. (1953). *Theory of games and economic behavior.* Princeton University press.

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

Olson, Eric T., "Personal Identity", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), retrieved from https://plato.stanford.edu/archives/fall2019/entries/identity-personal/

O'Malley-James, J. T., Cockell, C. S., Greaves, J. S. and Raven, J. A. (2014). Swansong biospheres II: The final signs of life on terrestrial planets near the end of their habitable lifetimes. *International Journal of Astrobiology, 13*(3), 229–243.

Omohundro, S. M. (2008). The basic AI drives. In *AGI* (pp. 483–492).

Ord, T. (2015). Moral Trade. *Ethics, 126*(1), 118–138.

Pamlin, D., & Armstrong, S. (2016). *Global Catastrophic Risks 2016.*

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

Parfit, D. (2012). Another Defence of the Priority View. *Utilitas, 24*(03), 399–440.

Pearce, D. (1995). *The Hedonistic Imperative.* Retrieved from: https://www.hedweb.com/hedethic/tabconhi.htm

Plant, M., & Kaczmarek, P. (Unpublished). This Time It's Personal: Incorporating 'Making People Happy, Not Making Happy People' into the Expected Moral Value Model.

Reese, J. (2018). Why I prioritize moral circle expansion over artificial intelligence alignment. *Effective Altruism Forum.* Retrieved from https://forum.effectivealtruism.org/posts/BY8gXSpGijypbGitT/why-i-prioritize-moral-circle-expansion-over-artificial

Rowe, T. and Beard, S. (2018). *Probabilities, methodologies and the evidence base in existential risk assessments.* Working paper. Centre for the Study of Existential Risk, Cambridge, UK.

Sandberg, A., & Bostrom, N. (2008). Global catastrophic risks survey. *Civil Wars, 98*(30), 4.

Sepielli, A. (2013). Moral uncertainty and the principle of equity among moral theories. *Philosophy and Phenomenological Research, 86*(3), 580-589.

Shulman, C. (2012). Are pain and pleasure equally energy-efficient? Retrieved from reflectivedisequilibrium.blogspot.com/2012/03/are-pain-and-pleasure-equally-energy

Singer, P. (1981). *The expanding circle.* Oxford: Clarendon Press.

Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica (Slovenia), 41*(4), 389–400.

Tainter, J. A. (1995). Sustainability of complex societies. *Futures, 27*(4), 397-407.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* Knopf.

Tonn, B., & MacGregor, D. (2009). A singular chain of events. *Futures, 41*(10), 706–714.

Torres, P. (2018). Space colonization and suffering risks: Reassessing the "maxipok rule." *Futures, 100*(April), 74–85.

Torres, P., & Rees, M. J. (2017). *Morality, foresight, and human flourishing: An introduction to existential risks.* Pitchstone Publishing.

Wright, R. (2001). *Nonzero: The logic of human destiny.* Vintage.