The expected value of the long-term future*

Thomas M. Sittler†

1 Introduction

A number of ambitious arguments have recently been proposed about the moral importance of the long-term future of humanity, on the scale of millions and billions of years. Several people have advanced arguments for a cluster of related views. Authors have variously claimed that shaping the trajectory along which our descendants develop over the very long run (Beckstead, 2013), or reducing extinction risk, or minimising existential risk (Bostrom, 2002), or reducing risks of severe suffering in the long-term future (Althaus and Gloor, 2016) are of huge or overwhelming importance. In this paper, I develop a simple model of the value of the long-term future, from a totalist, consequentialist, and welfarist (but not necessarily utilitarian) point of view. I show how the various claims can be expressed within the model, clarifying under which conditions the long-term becomes overwhelmingly important, and drawing tentative policy implications.

Views of the long term Beckstead (2013) defends the *long-run importance thesis*:

> 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.

Bostrom (2002), who accepts some version of the long-run importance thesis, defines an existential risk as one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential. He argues that reducing existential risks ought to be our foremost priority and proposes the *Maxipok* rule of thumb for prioritising altruistic actions:

> *Maximize the probability of an okay outcome*, where an "okay outcome" is any outcome that avoids existential disaster.

*This version: Author's manuscript, 2017-12-30 19:10:26+01:00

†I am grateful to Jan Brauner, Ryan Carey, Max Dalton, Richard Ngo and Brian Tomasik for helpful comments on earlier versions of this article

Beckstead, on the other hand, distinguishes between three ways we could shape the long-term future: (i) ripple effects of ordinary actions, (ii) existential risk reduction, and (iii) trajectory changes; he does not come down strongly in favour of any one of them.

Finally, Althaus and Gloor (2016) warn that the long-term future may be used for the creation of disvalue as well as value, stating that continued human development could also end up producing astronomical quantities of suffering. They denote such cases as "suffering risks" or "s-risks", arguing that these could constitute a priority from the perspectives of many value systems.

2 A basic model of the long-term future

In possible world w, the value of the long-term future is:

$$V_w = \int_{t=1}^{\infty} N_w(t) Q_w(t)$$

where $N_w(t)$ is the number of morally relevant beings at time t, and $Q_w(t)$ is the mean moral value of their lives at that time. $N_w(t)$ is understood to drop to 0 forevermore once we go extinct. Unfortunately, this expression is not very tractable, either intuitively or mathematically, so I change the the model to discrete time and add two important simplifying assumptions. First, I assume that if extinction in world w happens in period k, $N_w(t)$ is independent of k for $t < k$. In other words, population does not depend on whether extinction is imminent or far off. Second, I assume that $Q_w(t)$ and $N_w(t)$ are independent across possible worlds. These assumptions won't be important until sections 4 and 5, respectively, where I will discuss them again. With these assumptions, we get the *basic model* of equation 1. It says that the expected value of the long-term future is the sum over time of the product of three factors. The first is the probability $P(t)$ of reaching time t. The second and third are, conditional on reaching t, the expected number $N(t)$ of relevant beings and the expected mean moral value $Q(t)$ of their lives at that time. $Q(t)$ is the value of a life during one discrete interval, for example the value of a century-long life.

$$V = \sum_{t=1}^{\infty} P(t) N(t) Q(t) \tag{1}$$

The phrase "existential risk", which has recently become popular, packs many substantive assumptions into a single term. If we merely look at the basic model and make no additional empirical assumptions, we see that "existential risks" does not distinguish between:

- risks of extinction (low values of $P(t)$)
- permanent and drastic curtailment of potential, including
    - low or even negative values of $Q(t)$, for example because of a permanently stable repressive totalitarian global regime (Bostrom 2002).

- relatively low values of $N(t)$, for instance from a failure to flourish into a space-faring civilisation.

In fact, "existential risk" has sometimes been used interchangeably with "risks of extinction", omitting any reference to the future's quality or size (Althaus and Gloor, 2016). In the next three sections, I use the basic model to carefully distinguish the effects of changes in each of the three parameters.

3 $P(t)$ and extinction risk reduction

In this section, I draw upon the unpublished manuscript of *Modelling Risk Reduction* (Ord, 2014). Say that each time period has some probability r_i of going extinct, conditional on surviving all previous periods. We are now at the beginning of period 1. The probability of reaching the end of period t is then:

$$P(t) = \prod_{i=1}^{t}(1 - r_i) \tag{2}$$

It would be convenient to choose a period length for which extinction risk is not too small to intuitively think about, for example centuries. Depending on which parameters we let vary, we might then consider a number of different models.

3.1 Constant risk, temporary effects

Here, we take a version of equation 2, but we constrain the risk in every century after the first to be constant and equal to r.

$$\begin{aligned} P(t) &= (1 - r_1)\prod_{i=2}^{t}(1 - r) \\ &= (1 - r_1)(1 - r)^{t-1} \end{aligned}$$

We can only affect r_1, and

$$-\frac{dP(t)}{dr_1} = (1 - r)^{t-1}$$

is the value of reducing it by one unit.

Here, the lower the future risk r, the larger the value of reducing r_1. (A general point we will encounter throughout this paper is that the lower the aggregate risk after period t, the larger the value of reducing the risk up to period t). This is because, the lower the future risk, the longer the duration that is at stake. With our assumptions, the expected duration $D = \sum_{t=1}^{\infty} P(t)$ of our civilisation is a geometric series, and converges to

$$D = \frac{1 - r_1}{r}$$

If the risk per century was 50%, the expected duration would be one century ($r_1 = r = 0.5$, $D = 1$). If it were 1%, then the expected duration would be 99 centuries ($r_1 = r = 0.01$, $D = 99$). Furthermore, if $r_i = r$, half of the expected duration comes from possible worlds where civilisation lasts for $-\ln(2)/ln(1-r)$ centuries or less, 69 centuries in the case of $r = 0.01$. More generally a proportion x of the duration comes from futures of length $\ln(1-x)/\ln(1-r)$ or less.[1] Unless r is extremely low, on this model the long-term future isn't really about the next billions of years, but rather the next hundreds of thousands or at most millions of years (see section 5.1). To what degree could we affect D? $-\frac{dD}{dr_1} = \frac{1}{r}$, so if $r = 0.5$, then reducing risk this century by one percentage point would increase the expected duration by 2 years; if $r = 0.01$, then reducing the risk this century by one percentage point would increase the expected duration by one century. Hence it would appear that if future risk is low, then reducing the risk this century could easily be our best option for producing altruistic value.

3.1.1 Diminishing returns on risk reduction Unfortunately,

> while lower levels of extinction risk make eliminating each percentage point of risk more valuable, we should also expect that it becomes much harder to eliminate a percentage point of risk when very few remain. A plausible model would be that it is roughly as difficult to halve the risk per century, regardless of its starting probability, and more generally, that it is equally difficult to reduce it by some proportion regardless of its absolute value beforehand. (Ord, 2014)

To handle diminishing returns, it's natural to make two changes to equation 2. First, let f_1 be the fraction by which we will reduce existing risk r_1 this century, giving us

$$P(t) = (1 - r_1(1 - f_1)) \prod_{i=2}^{t} (1 - r_i) \tag{3}$$

Second, we again suppose that the risk after this century is constant, and we additionally let it be equal to this century's pre-intervention risk r_1. Hence r_i are constant, and we get $P(t) = (1 - r(1 - f_1)) \cdot (1 - r)^{t-1}$, and the expected duration is

$$D = \frac{1 - r(1 - f_1)}{r} = \frac{1}{r} - 1 + f_1$$

When we evaluated $-\frac{dD}{dr_1}$ above, a motivating assumption was that for one unit of resources, we could buy a one percentage point decrease in r_1. But we

[1] Futures of length L or less contribute $\sum_{t=1}^{L} P(t) = \frac{(1-(1-r)^L)(1-r)}{r}$ to D, so the fraction contributed by futures of length L or less is $(1 - (1 - r)^L)$. Solving $x = (1 - (1 - r)^L)$ for L gives $L = \frac{\ln(1-x)}{\ln(1-r)}$.

have found this constant returns assumption lacking. When we say instead that it is equally difficult to reduce extinction risk by some proportion regardless of its absolute value beforehand, we mean that one unit of resources purchases a one percentage point increase in f_1 (recall that f_1 is defined as the fraction by which we *reduce* existing extinction risk). Hence we are interested in $\frac{dD}{df_1} = 1$, which has no dependence on r. As Ord (2014) writes:

> The two effects of small values of r (increasing the expected value of the future and making it harder to get a given level of absolute risk reduction) exactly cancel. It turns out on this model that even if we could reduce the level of existential risk[2] in the next century by half, the expected value of this would just be half the value of a century of humanity. If we were to approach the complete elimination of the risk in the next century, the expected value achieved would just approach the full value of a century of humanity.
>
> As we said above, this is a very large, but not overwhelming, amount of value. Its scale is not out of keeping with how difficult it would be to achieve. For example, in order to reduce existential risk over the century by half, we may have to give up new technologies or live with much stronger security or surveillance measures.

3.1.2 The trivial model There is one simpler model which is often used informally in justifications of the importance of extinction risk reduction. Following Ord, I call the trivial model that which multiplies A, the value of humanity reaching its full potential, with one minus the total amount of extinction risk R to obtain the value of the future:

$$V = (1 - R)A$$

Since A is often judged to be immense, the value of reducing extinction risk by even a small amount could easily trump all other altruistic priorities.

The trivial model is unhelpful because it's very difficult to get an intuitive grasp on R or on A. Surviving for only 1000 centuries before extinction may not mean we have reached our full potential, but it clearly is better than nothing. A makes no allowance for such partial successes. Even if we set aside this problem by assuming that success is a very binary thing (we either do or don't reach a risk-free utopia), it remains difficult to estimate R, which is the entire risk of failing to reach utopia. And it is still harder to estimate how large an impact we could have on R. At worst, we might slip into equating R with r_1, the risk this century.

So far I have shown that the trivial model has severe shortcomings, relative to my basic model. Coupling the basic model with the fairly natural assumption of diminishing returns drastically reduces the expected value of extinction risk reduction. As Ord (2014) writes, "there is either not that much future to come

[2]Ord uses existential risk, but for reasons I discuss above, I prefer to talk only of extinction risk.

(if risk per century is high), or we can't make a large absolute change in the chance of making that future happen (if the risk per century is low). Moreover, there is no sweet spot between these extremes."

There are other ways, however, of maintaining the view that extinction risk reduction is our foremost altruistic priority. One way, discussed in section 3.2, is to say that, conditional on surviving some small number of centuries, the risk in every century thereafter is very low. Another approach (section 3.3) is to argue that we have some actions available that would reduce the risk in all or many time periods at once.

3.2 Variable risk, temporary effects

3.2.1 How plausible is the assumption of constant risk? The constant risk model, if r is not negligible, assigns extremely low probabilities to distant periods. Beckstead (2013) argues that this is overconfident:

> Given the great uncertainty involved, including uncertainty about what people will do to prepare for these risks, it would seem overconfident to have a very high probability or a very low probability that humans will survive for [a] full billion years. Having a very high or low probability in this claim, such as less than 1% or greater than 99%, would require much greater certainty about the future than it is reasonable to have. Obviously, choosing any specific number here would be arbitrary. To be conservative, I will assume that our subjective probability in this claim should be at least 1%.

Whether a 1% chance of surviving for a billion years seems conservative or daring is not, I argue, a decisive consideration one way or another. To get closer to the crux of the issue, we need to introduce explicit empirical assumptions about what the causes of extinction risk will be.

On one simple picture of extinction risk, a small number of causes (for example, major-power wars and pandemics) are responsible for most of the risk humanity faces (the *few causes* view). In such a world, constant risk would be quite implausible. Recall that r_i is the risk in period i, *conditional* on surviving up to that point. Suppose that from today, humanity survived for another thousand centuries. We would then have compelling evidence that the underlying causes of risk have been effectively neutralised. Had they not, we would have been unlikely to survive for a thousand centuries. Hence more generally if there are a small number of causes of risk, only decreasing r_i are plausible. This is true even if the risks are extremely dangerous and make the unconditional probability of human survival low; conditional on surviving a sufficiently long time, it's likely we will continue to survive.

A picture which makes constant risk more plausible is the following. We may think of human history as the process of extracting balls from a giant urn of possible technologies (Bostrom, 2013). So far, all the balls we have extracted from this urn have been white or grey, meaning that they have been beneficial,

or perhaps mixed blessings. We have not so far pulled out a black ball from this urn — we have not made a discovery that would be highly likely to spell the end of our civilisation. If the number of possible technologies is in practice unlimited (the *bottomless urn* picture), then even in the scenarios where we keep surviving draw after draw from the urn, there should remain some independent risk from new draws. In other words, as humanity progresses, it keeps discovering ever more powerful technologies, each of which may spell doom even if we have learnt to safely use every previous technology. Even if this independent risk were only one in a million per century, surviving for a billion years –ten million centuries– would be virtually impossible. If we re-frame the proposition that we will survive for a billion years as a conjunction of ten million sub-propositions (we'll survive century 1 and century 2 and ...), it no longer seems intuitively so overconfident to assign it a very low probability.

Anthropogenic risks (like nuclear weapons) are generally thought to loom larger than natural ones (like asteroids). If this is true, the most natural way to lend credence to the few causes view is with the scenario of *technological maturity*, a state where we have developed "all the major technologies that are feasible and have survived their creation and utilisation by society" (Ord, 2014). If eventual technological maturity is likely conditional on survival, then to say that we will survive for a billion years (or much longer) is to say that we will survive the discovery of every feasible technology — a relatively shorter conjunction. However, it could still be plausible that hundreds of independently risky technologies are required to reach maturity. If the independent risk from 500 of these technologies were 1% each, the probability of reaching maturity would be only $0.99^{500} \approx 0.6\%$.

Other than technological maturity, a second scenario for decreasing risk is one we might call *risk-independent islands*. Here, humanity colonises space, perhaps establishing permanent settlements around a number of stars. Then "local" risks, whose probabilities are not correlated across locations, are no longer a threat to the entire future. Instead, the risk rate would be lowered to those risks which could realistically affect our entire region of space (Ord, 2014).

3.2.2 A model for decreasing risk The same point may be made more generally, for an event E, instead of specific scenarios. Suppose again that we can only affect f_1, and that the risk is r_α from period 1 until an event E at the end of period j which lowers the risk to r_Ω forevermore. Then for $t > j$,

$$P(t) = (1 - r_\alpha(1 - f_1))(1 - r_\alpha)^{j-1}(1 - r_\Omega)^{t-j} \tag{4}$$

If r_Ω were sufficiently small, and E not to far off, this would suffice to make D very large, *even if* we assume diminishing returns on our ability to affect r_1.[3]

[3]In general $D = \frac{(1-r_\alpha(1-f_1))(1-(1-r_\alpha)^j)}{r_\alpha} + \frac{(1-r_\alpha(1-f_1))(1-r_\alpha)^{j-1}\cdot(1-r_\Omega)^{j+1}}{r_\Omega}$, where the first fraction is the contribution to D of periods 1 to j, and the second fraction is the contribution to D of periods $j+1$ to infinity.

For example, if a thousand years hence the risk dropped from 10% to 0.1%, ($r_\alpha = 0.1$, $r_\Omega = 0.001$ $j = 10$) we would get $D \approx 351 + 39f_1$, meaning that halving the risk this century would increase D by 19.5 centuries; and if we were to approach the complete elimination of the risk this century, the expected duration increase achieved would approach 39 centuries.

3.2.3 Increasing risk We might also consider a model where the risks increase over time. If natural risks are low, the risk today, with nuclear weapons, is likely higher than it has been at least since the development of agriculture. Perhaps as technology progresses, every civilisation reaches a region of the urn containing many black balls. For example, perhaps we will inevitably develop some hypothetical weapons that give so large an advantage to offence over defence that civilisation is certain to be destroyed. This sort of scenario, akin to a "great filter" proposed by Hanson (1998), would be a compelling explanation for the Fermi paradox. We could model this by re-purposing equation 4 and letting r_Ω be larger than r_α instead of smaller. The expected duration of civilisation would then depend on how much time we have left before the great filter.

3.3 Constant risk, lasting effects

Suppose we were able to take some actions that affect the risk in all time periods. Ord (2014) argues that such measures plausibly exist:

> Part of surviving the rise of nuclear weapons involved developments in arms control, which might be transferable to some other future risks. The foundational work about the concept of existential risk should help us to prioritise them and should generalise across risks. If we could develop stable institutional structures for addressing risks as they arise, these could also provide value across the future. In general, it seems that some risk reduction measures reduce risks in all centuries and some just reduce the risks in that century.

Recalling equation 3, $P(t) = (1-r_1(1-f_1))\prod_{i=2}^{t}(1-r_i)$, we now generalise it by letting f_i equal the amount by which we will reduce the risk in period i. Assuming again, to keep the model tractable, that the pre-intervention risk r is constant, we get $P(t) = \prod_{i=1}^{t}(1-r(1-f_i)))$. Assuming that the fraction by which any generation in fact reduces the risk is independent of time, we can say $f_i = f$. Hence

$$P(t) = (1-r(1-f))^t = (1-r+rf)^t$$

and $D = \frac{1-r+rf}{r-rf}$, and hence, $\frac{dD}{df} = \frac{1}{(f-1)^2 r}$. This grows large as r diminishes, again allowing the value of risk reduction to become overwhelming if r is very small, or modest if r is large. We may also note that the effect of f on D is now quadratic in f.

Is $f_i = f$ realistic? Ord argues that we may be able to expect future generations to be more interested in risk reduction, implying increasing f_i:

> If existential risk reduction turns out to be clearly worth doing, then we might expect that people may eventually become convinced of this and start systematically reducing it. If so, there might be much more risk in the coming century or two than in the later centuries where people begin addressing it regardless of our actions now.

This would only reinforce the case for extinction risk reduction. Future generations systematically reducing risk would in effect be a kind of event E.

4 $N(t)$ and becoming a space-faring civilisation

The future is often said to have overwhelming expected value because, in addition to being potentially very long, it is potentially very populous. It is often thought likely that humanity, conditional on surviving, will eventually conquer the stars, leading to astronomical population sizes. Bostrom (2003) writes that "the Virgo Supercluster could contain 10^{23} biological humans".

Recall that in section 2 we posited that if extinction in world w happens in period k, $N_w(t)$ is independent of k for $t < k$. Note that none of what we have said so far, about $P(t)$ and D, depends on this assumption. But the assumption does become important when we multiply $P(t)$ and $N(t)$. Complete independence is obviously unrealistic: it would be silly to assume that there could not even be a small correlation. However, multiplying $P(t)$ with $N(t)$ could still be a useful modelling assumption if the correlation is not too large. In general I find approximate independence plausible. There is one case in which it clearly is not: if space colonisation is in fact likely to involve risk-independent islands. Then high population goes with low risk, increasing the value of the future relative to the basic model.

Continuing now with the independence assumption, the relevance of the expected population $N(t)$ to the argument for treating the long-term future as overwhelmingly important depends on our modelling assumptions in a number of ways.

If we think the risk is decreasing sufficiently quickly (section 3.2) constant $N(t)$ is sufficient to make the argument work. If we think the risk is likely to remain relatively large, and we continue with the natural assumption of diminishing returns, $N(t)$ does play an important role. Suppose we return to equation 3, and let r_i be constant for $i > 1$, but $N(t)$ grows exponentially by a factor p from a base of $N(1)$. The expected number of people who will live in period t is

$$\begin{aligned} P(t)N(t) &= (1 - r_1(1 - f_1)) \cdot (1 - r)^{t-1} \cdot N(1)(1 + p)^t \\ &= (1 - r_1(1 - f_1)) \cdot (1 - r)^{-1}[(1 - r)(1 + p)]^t N(1) \end{aligned} \tag{5}$$

If the growth rate of population exactly cancels out the risk of extinction ($(1-r)(1+p) = 1$), then the effect of f_1 on the expected number of people who

will live in total over the next n periods is

$$\frac{d\sum_{t=1}^{n} P(t)N(t)}{df_1} = n\frac{r_1 N(1)}{1-r}$$

allowing the expected value of increasing f_1 to be very large.[4]f

Taking humans as the population of interest, it's clear that $(1-r)(1+p)$ has recently been even greater than one. Even on an alarming prediction, such as $r = 0.5$, population would only have to double each century in order to make $(1-r)(1+p) = 1$. In fact, world population has more than tripled over the last century. However, this increase in population may be due to stop soon, and it clearly cannot go on for many centuries unless we become a space-faring civilisation.

If we were less pessimistic about the risk in the next few centuries, say $r = 0.1$, we would only need $p = 1/9 \approx 11\%$ per century to make $(1-r)(1+p) = 1$. If this rate of population were to go on for 1500 years hence, our population would only increase by a factor of $\frac{10}{9}^{15} \approx 4.9$ relative to today, which could, be sustainable even just on earth. On this model (and supposing $r_1 = r$)

$$\frac{d\sum_{t=1}^{15} P(t)N(t)}{df_1} = 15 \cdot \frac{10}{9}N(1)$$

allowing us to create *at least* the equivalent of about $\frac{25}{3}$ centuries of civilisation at current population levels (or 6.3 trillion life-years) by halving the risk this century. This amount of value is very large, but not astronomical, and may or may not be our best altruistic option. This shows that if we believe the risk over the next few centuries won't be too high, eventual space-travel need not be posited to make reducing extinction risk a top priority. With higher risk, space travel may be necessary. To say exactly how population growth from space travel interacts with high extinction risk estimates is straightforward to do using equation 5, but the mathematics becomes more tedious.

We have seen in this section that an increase in $N(t)$ can cancel out a decrease in $P(t)$ — what the relative sizes of the marginal increases must be depends on the previous levels of $N(t)$ and $P(t)$. This means it's worth looking at which of the two might be more tractable. The forces driving human demography while on earth are likely to be very difficult to push against. The likelihood of space colonisation, a high-profile issue on which billions of dollars is spent per year (Masters, 2015), also seems relatively hard to affect. Extinction risk reduction, on the other hand, is relatively neglected (Bostrom, 2013; Todd, 2017), so it could be easier to achieve progress in this area.

[4]In fact, it is unbounded on this model. But indefinite exponential growth is prohibited by our current understanding of physics. In the long run, the best we can hope for is to grow cubically with time, as our civilisation expands outwards into the cosmos like a sphere. We might model this with $N(t) = N(s)(ct)^3$ for $t \geq s$, where s is the century in which we begin to colonise space and c is some constant. Since exponentials always beat polynomials in the limit of t, this will allow $\sum_{t=1}^{\infty} P(t)N(t)$ to remain finite.

5 $Q(t)$ and future flourishing

It's first worth noting that, although reducing risks of extinction has been the primary focus of many of those who believe the long-run importance thesis, the thesis does not alone imply that risk reduction is the top priority. If we could cause increases in $Q(t)$ that are sufficiently long-lasting, there is nothing in the basic model that suggests this would be less effective than increasing $P(t)$. For example, suppose that some aspect of $Q(t)$'s future trajectory could be path-dependent. Events that shift us to a different path may be called *trajectory changes*. A plausible example would be the entrenchment of political systems or values that become very impervious to change. Positive trajectory changes could be competitive with extinction risk reduction.

Before I delve into further discussion of $Q(t)$, I return, as promised, to the second independence assumption, which becomes relevant when we evaluate V. Recall that in section 2, we assumed that $Q_w(t)$ and $N_w(t)$ are independent across possible worlds w. $N_w(t)$ is likely to be correlated with the degree of technological sophistication of our civilisation. Moreover, to the extent that advanced technologies can be used to make one's own life very valuable (a likely but not foregone conclusion), those who control these technologies will contribute to a high $Q_w(t)$. Hence the more one believes that technology will make life valuable, and that a large majority of future sentient beings will have access to this technology (see section 5.2.2), the more correlated $N_w(t)$ and $Q_w(t)$.

5.1 Time-scales

Which are the values of t that matter most? This depends on one's views about the shape of $P(t)$ and $N(t)$. To the total number $\sum_{t=1}^{\infty} P(t)N(t)$ of people who are expected to ever live, how much is contributed by possible worlds where civilisation goes on for very long (long futures) and how much is contributed by shorter futures? This depends on our model parameters. For example, if we think the risk will eventually decrease to a very low value (see section 3.2), then most of the value comes from long futures. If risk and population are both constant, a proportion x of the value comes from futures of length $\ln(1-x)/\ln(1-r)$ or less. More general expressions, in terms of $N(t)$, can be found using the model.[5] Moreover, since we are uncertain about which model is the correct one, we should focus our efforts, likelihood being equal, on improving those futures which are very large. If we have sufficient belief in models giving astronomically large futures, we should focus only on these.

[5] For example, if $N(t) = N(1)(ct)^3$, the proportion x contributed by futures of length L or less is $x = -((r-1)(L^3r^3(1-r)^L + 3L^2r^2(1-r)^L - 3Lr^2(1-r)^L + r^2(1-r)^L + 6Lr(1-r)^L - 6r(1-r)^L + 6(1-r)^L - r^2 + 6r - 6))\frac{1}{r^3-7r^2+12r-6}$. Solving this for L is left as an exercise to the reader.

5.2 The possibility of a bad future

Our discussion so far has been based on the implicit premise that the future will, on balance, be good, or, in welfarist terms, that the average well-being $Q(t)$ will be a positive number, at least for the values of t that matter. We might call this view *future-optimism*. If most of the future were bad (*future-pessimism*), increasing $N(t)$ and $P(t)$ would generally do more harm than good.

For some combination of the assumptions discussed in section 5.1, we can essentially ignore small futures, which means ignoring short futures. Some people believe that it's nearly impossible to have a consistent impact on $Q(t)$ so very distant futures. If this is true, we find ourselves in the unenviable position of being forced to form our best guess about whether the future will be good or bad, and then increase or reduce the probability of extinction accordingly. We would have to make a decision with astronomical stakes based on very little evidence.

In the next two sections, I discuss possible sources of future suffering, with the proviso that my speculations are all the more uncertain the further they are extrapolated into the future.

5.2.1 Suffering from conflict One big category of suffering, about which I don't have much to add, are wars and other conflicts. Could there be a state of conflict for a significant portion of the long-run future? For this to be possible, there need to be two or more sets of values fighting each other with violence for long periods of time. In a two-player war of attrition game, the mixed strategy Nash equilibrium is such that the players' expected pay-off is zero; in expectation, the entire value of the prize will be wasted in war. In general, lasting wars are bad for everyone involved, who would be better off conducting trade, both of the commercial and the moral kind (Tomasik, 2013; Ord, 2015). So they are likely to be caused by failures of coordination. This argument only yields the familiar conclusion that cooperation and geopolitical stability are to be promoted, though perhaps even more so than we might have thought before considering the long-term future.

5.2.2 Suffering of powerless persons Leaving aside conflict between sets of values, in this section we will ask two questions, both essentially related to which values will control the future. First, will there be powerless sentient beings in the future? By this I mean sentients who don't control their own lives to a sufficient extent. Second, will the beings (or processes, or institutions) that control the future be impartial altruists, or will they be selfish? Of course, there are really degrees of selfishness and altruism, I focus on the extremes only to focus our thinking.

A plausible hypothesis is that those who control the future will have good lives. Suppose that life in the future were to become not worth living, and inescapably so. Would those who are in charge at that time be able to make it stop, either by committing suicide or at least by ensuring they will not have descendants? Today, a very determined and reasonably enterprising individual

has access to painless forms of suicide, such as carbon monoxide poisoning, or to sterilisation. However, this individual may be biased against ending things, for instance because of the survival instinct, and so could individuals or groups in the future. The extent of this bias is an open question. Overall, it seems plausible to me that those who control the future retain the "option value" to bring about extinction. If this is true, net-negative lives in the future would have to be the lives of those of its inhabitants who don't control it, and don't enjoy option value.

We may further speculate that if the future is controlled by altruistic values, even powerless persons are likely to have lives worth living. If society is highly knowledgeable and technologically sophisticated, and decisions are made altruistically, it's plausible that many sources of suffering would eventually be removed, and no new ones created unnecessarily. Selfish values, on the other hand, do not care about the suffering of powerless sentients.

What could be the scale of such suffering? By way of illustration, we may look at our own past. Over the majority of our history, the suffering of powerless humans has been tremendous. Pinker (2011) documents in detail the scale of past cruelty and indifference, including the subjection of women, the persecution of minority groups, the brutal torture of petty criminals, witch-burning, and so on. Children, who were widely considered less than human, did not enjoy legal protection against physical maltreatment in the United States until the latter half of the nineteenth century (Pinker, 2011, chapter 7). While our moral circle now generally extends to all members of our own species, we still, systematically and at an industrial scale, inflict suffering on non-human animals. The Cambridge declaration on conciousness (Low et al., 2012) affirms the scientific consensus that non-human animals are sentient. Yet 65 billion chickens were killed for meat in 2016 (Food and Agriculture Organization of the United Nations). Most of these animals are selectively bred for extremely rapid growth, which leads to lifelong severe suffering from poor bone health and leg disorders including deformities, lameness, tibial dyschondroplasia, and ruptured tendons (The Humane Society of the United States, 2013). Each year 80 billion farmed fish are killed (Mood and Brooke, 2012), and 1-3 trillion wild fish are caught (Mood and Brooke, 2010) and killed mostly by asphyxiation. If, in a selfish future, the number of animals whose suffering we directly cause continues to be several times the number of humans, anyone who gives some significant moral weight to non-human animals may be forced to conclude that the continued survival of our civilisation would be an evil.[6] The case of non-human animals is just an illustration. More generally, future society could be organised in any

[6] When we consider Darwinian suffering (Ng, 1995), which is not directly our doing, our indifference is even greater. Wild animals, who number in the trillions (Tomasik, 2009), routinely experience intense suffering from predation, starvation and disease. Unlike with farmed animals, we couldn't substantially reduce this suffering by simply changing our consumption patterns, so the evidence it provides about our values is less direct. Moreover, the direction of the effect of continued human civilisation on wild animal welfare is unclear. If wild animals have lives not worth living (Tomasik, 2015; Sittler-Adamczewski, 2016), civilisation is a negative if we spread wild ecosystems to the stars, but a likely positive if we do not. If wild animals enjoy net-positive lives, the reverse is true.

number of ways that bad for the lives of powerless sentients, including ways that are much worse than the current situation.

Selfishness is not a sufficient condition for large-scale future suffering. In today's world, we rear chickens in factory farms, because we are, by and large, sufficiently selfish to do so, but also because it is an economically efficient method of meat and egg production. It's an open question to what extent it will be economically efficient for future selfish values to make or let powerless beings suffer.

In this discussion, there are two considerations that might at first have appeared to be crucial, but turn out to look less important. The first such consideration is whether existence is in general good or bad, à la Benatar (2008). If existence really should turn out to be a harm, sufficiently unbiased descendants would plausibly be able to end it. This is the option value argument. In turn, option value itself might appear to be a decisive argument against doing something so irreversible as ending humanity: we should temporise, and delegate this decision to our descendants. But not everyone enjoys option value, and those who suffer are relatively less likely to do so. If our descendants are selfish, and find it advantageous to allow the suffering of powerless beings, we may not wish to give them option value. If our descendants are altruistic, we do want civilisation to continue, but for reasons that are more general than option value.

5.2.3 What should future pessimists do? Future pessimists, more than we might at first have supposed, have some good reasons against attempting to destroy the world. Doing so would strongly violate the preferences of many people, with whom pessimists may be better off cooperating than fighting. If one side is trying to increase risks of extinction, and the other side is trying to do the opposite, some portion of each side's efforts cancels out. Both can then benefit from moral trade where each redirects that portion towards their shared goal of increasing $Q(t)$. In other words, pessimists could offer a compromise: We'll let you spread into the cosmos if you give more weight to our concerns about future suffering (Tomasik, 2013).

One way to increase $Q(t)$ without substantially affecting $P(t)$ in either direction is to advocate for positive value changes in the direction of greater consideration for powerless sentients, or to promote moral enhancement (Persson and Savulescu, 2008). Another approach might be to work to improve political stability and coordination, making long-term conflict less likely as well as increasing the chance that moral progress continues.

6 Conclusion

I have shown how moving on from the trivial model that is sometimes implicitly used can help us better understand which assumptions are needed to make the long-term future overwhelmingly important; and I have discussed the plausibility of a bad future, which would dramatically reverse our conclusions about the value of extinction risk reduction.

Suppose we accept some combination of views leading to the long-run importance thesis. Then, we ought to do whatever most effectively increases the value of the long-term future. How should we do this? This question, of course, is an entire research agenda in itself. Even within our basic model, a full answer to this question would quickly become intractable, as it depends on so many model parameters. Nonetheless, our discussion of the basic model has suggested a few tentative hypotheses.

Contents

1 Introduction **1**

2 A basic model of the long-term future **2**

3 $P(t)$ and extinction risk reduction **3**
3.1 Constant risk, temporary effects 3
3.1.1 Diminishing returns on risk reduction 4
3.1.2 The trivial model 5
3.2 Variable risk, temporary effects 6
3.2.1 How plausible is the assumption of constant risk? 6
3.2.2 A model for decreasing risk 7
3.2.3 Increasing risk 8
3.3 Constant risk, lasting effects 8

4 $N(t)$ and becoming a space-faring civilisation **9**

5 $Q(t)$ and future flourishing **11**
5.1 Time-scales 11
5.2 The possibility of a bad future 12
5.2.1 Suffering from conflict 12
5.2.2 Suffering of powerless persons 12
5.2.3 What should future pessimists do? 14

6 Conclusion **14**

References

Althaus, D. and Gloor, L. (2016). Reducing Risks of Astronomical Suffering: A Neglected Priority.

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future*. PhD thesis, Rutgers University - Graduate School - New Brunswick.

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(1).

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

Bostrom, N. (2013). Existential Risk Prevention as Global Priority: Existential Risk Prevention as Global Priority. *Global Policy*, 4(1):15–31.

Food and Agriculture Organization of the United Nations. FAOSTAT database. http://www.fao.org/faostat/.

Hanson, R. (1998). The Great Filter. http://mason.gmu.edu/˜rhanson/greatfilter.html.

Low, P., Panksepp, J., Reiss, D., Edelman, D., Van Swinderen, B., and Koch, C. (2012). The Cambridge declaration on consciousness. In *Francis Crick Memorial Conference, Cambridge, England*.

Masters, K. (2015). How much money is spent on space exploration? http://curious.astro.cornell.edu/about-us/150-people-in-astronomy/space-exploration-and-astronauts/general-questions/921-how-much-money-is-spent-on-space-exploration-intermediate.

Mood, A. and Brooke, P. (2010). Estimating the Number of Fish Caught in Global Fishing Each Year.

Mood, A. and Brooke, P. (2012). Estimating the Number of Farmed Fish Killed in Global Aquaculture Each Year.

Ng, Y.-K. (1995). Towards welfare biology: Evolutionary economics of animal consciousness and suffering. *Biology and Philosophy*, 10(3):255–285.

Ord, T. (2014). Modelling Risk Reduction.

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

Persson, I. and Savulescu, J. (2008). The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity. *Journal of Applied Philosophy*, 25(3):162–177.

Pinker, S. (2011). *The Better Angels of Our Nature: The Decline of Violence In History And Its Causes*. Penguin UK.

Sittler-Adamczewski, T. M. (2016). Consistent Vegetarianism and the Suffering of Wild Animals. *Journal of Practical Ethics*, 4(2):94–102.

The Humane Society of the United States (2013). The Welfare of Animals in the Chicken Industry. Technical report.

Todd, B. (2017). The case for reducing extinction risks. https://80000hours.org/articles/extinction-risk/.

Tomasik, B. (2009). How Many Wild Animals Are There?

Tomasik, B. (2013). Gains from Trade through Compromise.

Tomasik, B. (2015). The Importance of Wild-Animal Suffering Wild Animal Suffering and Intervention in Nature: Studies and Research Contributions. *Relations: Beyond Anthropocentrism*, 3:133–152.