# Risks of Astronomical Future Suffering

Brian Tomasik
Center on Long-Term Risk
brian.tomasik@longtermrisk.org

### Abstract

It's far from clear that human values will shape an Earth-based space-colonization wave, but even if they do, it seems more likely that space colonization will increase total suffering rather than decrease it. That said, other people care a lot about humanity's survival and spread into the cosmos, so I think suffering reducers should let others pursue their spacefaring dreams in exchange for stronger safety measures against future suffering. In general, I encourage people to focus on making an intergalactic future more humane *if* it happens rather than making sure there will be an intergalactic future.

## Contents

## 1 Epigraphs

If we carry the green fire-brand from star to star, and ignite around each a conflagration of vitality, we can trigger a Universal metamorphosis. [...] Because of us [...] Slag will become soil, grass will sprout, flowers will bloom, and forests will spring up in once sterile places.[1][...] If we deny our awesome challenge; turn our backs on the living universe, and forsake our cosmic destiny, we will commit a crime of unutterable magnitude.
– Marshall T. Savage, *The Millennial Project: Colonizing the Galaxy in Eight Easy Steps*, 1994

Let's pray that the human race never escapes from Earth to spread its iniquity elsewhere.
– C.S. Lewis

If you can't beat 'em, join 'em.
– proverb

## 2 Humans values may not control the future

Nick Bostrom's "The Future of Human Evolution" (Bostrom, 2004) describes a scenario in which human values of fun, leisure, and relationships may be replaced by hyper-optimized agents that can better compete in the Darwinian race to control our future light cone. The only way we could avert this competitive scenario, Bostrom suggests, would be via a "singleton" (Bostrom, 2006), a unified agent or governing structure that could control evolution. Of course, even a singleton may not carry on human values. Many naive AI agents that humans might build may optimize an objective function that humans find pointless. Or even if humans do maintain hands on the steering wheel, it's far from guaranteed that we can preserve our goals in a stable way across major self-modifications going forward.

These factors suggest that even conditional on human technological progress continuing, the probability that human values are realized in the future may not be very large. Carrying out human values seems to require a singleton that's not a blind optimizer, that can stably preserve values, and that is shaped by designers who care about human values rather than selfish gain or something else. This is important to keep in mind when we imagine what future humans might be able to bring about with their technology.

Some people believe that sufficiently advanced superintelligences will discover the moral truth and hence necessarily do the right things. Thus, it's claimed, as long as humanity survives and grows more intelligent, the right things will eventually happen. There are two problems with this view. First, Occam's razor militates against the existence of a moral truth (whatever that's supposed to mean). Second, even if such moral truth existed, why should a superintelligence care about it? There are plenty of brilliant people on Earth today who eat meat. They know perfectly well the suffering that it causes, but their motivational systems aren't sufficiently engaged by the harm they're doing to farm animals. The same can be true for superintelligences. Indeed, arbitrary intelligences in mind-space needn't have even the slightest inklings of empathy for the suffering that sentients experience.

## 3 Some scenarios for future suffering

Even if humans do preserve control over the future of Earth-based life, there are still many ways in which space colonization would multiply suffering. Following are some of them.

### 3.1 Spread of wild animals

Humans may colonize other planets, spreading suffering-filled animal life via terraforming. Some humans may use their resources to seed life throughout the galaxy, which some sadly consider a moral imperative.

### 3.2 Sentient simulations

Given astronomical (Bostrom, 2003) computing power, post-humans may run various kinds of simulations. These sims may include many copies of wild-animal life, most of which dies painfully shortly after being born. For example, a superintelligence aiming to explore the distribution of extraterrestrials of different sorts might run vast numbers of simulations (Thiel, Bergmann and Grey, 2003) of evolution on various kinds of planets. Moreover, scientists might run even larger numbers of simulations of organisms-that-might-have-been, exploring

[1] Because nature contains such vast amounts of suffering, I would strongly dislike such a project. I include this quotation for rhetorical effect and to give a sense of how others see the situation.

the space of minds. They may simulate decillions of reinforcement learners that are sufficiently self-aware as to feel what we consider conscious pain.

### 3.3 Suffering subroutines

It could be that certain algorithms (say, reinforcement agents (Tomasik, 2014)) are very useful in performing complex machine-learning computations that need to be run at massive scale by advanced AI. These subroutines might be sufficiently similar to the pain programs in our own brains that we consider them to actually suffer. But profit and power may take precedence over pity, so these subroutines may be used widely throughout the AI's Matrioshka brains.

### 3.4 Black Swans

The range of scenarios that we can imagine is limited, and many more possibilities may emerge that we haven't thought of or maybe can't even comprehend.

## 4 Even a human-controlled future is likely to increase suffering

If I had to make an estimate now, I would give ~70% probability that if humans choose to colonize space, this will cause more suffering than it reduces on intrinsic grounds (ignoring compromise considerations discussed later). Think about how space colonization could plausibly reduce suffering. For most of those mechanisms, there seem to be counter-mechanisms that will *increase* suffering at least as much. The following sections parallel those above.

### 4.1 Spread of wild animals

David Pearce coined the phrase "cosmic rescue missions" (Pearce, n.d.) in referring to the possibility of sending probes to other planets to alleviate the wild extraterrestrial (ET) suffering they contain. This is a nice idea, but there are a few problems.

- We haven't found any ETs yet, so it's not obvious there are vast numbers of them waiting to be saved from Darwinian misery.
- The specific kind of conscious suffering known to Earth-bound animal life would not necessarily be found among the ETs. Most likely ETs would be bacteria, plants, etc., and even if they're intelligent, they might be intelligent in the way robots are without having emotions of the sort that we care very much about. (However, if they were very sophisticated, it would be relatively unlikely that we would not consider them conscious.)
- It's unclear whether humanity would support such missions. Environmentalists would ask us to leave ET habitats alone. Others wouldn't want to spend the energy on rescue missions unless they planned to mine resources from those planets.

Contrast this with the possibilities for *spreading* wild-animal suffering:

- Humans may spread life to many planets (e.g., Mars via terraforming, other Earth-like planets via directed panspermia). The number of planets that can support life may be appreciably bigger than the number that already have it. (See the discussion of $f_l$ in the Drake equation.) Moreover, the percentage of planets that can be converted into computers that could simulate wild-animal suffering might be close to 100%.
- We already know that Earth-based life is sentient, unlike for ETs.
- Spreading biological life is slow and difficult, but disbursing small life-producing capsules is easier than dispatching Hedonistic Imperative probes or berserker probes.

Fortunately, humans might not support spread of life that much, though some do. For terraforming, there are survival pressures to do it in the near term, but probably directed panspermia is a bigger problem in the long term. Also, given that terraforming is estimated to require at least thousands of years, while human-level digital intelligence should take at most a few hundred years to develop, terraforming may be a moot point from the perspective of catastrophic risks, since digital intelligence doesn't need terraformed planets.

While I noted that ETs are not guaranteed to be sentient, I do think it's moderately likely that consciousness is fairly convergent among intelligent civilizations. This is based on (a) suggestions of convergent consciousness among animals on Earth and (b) the general principle that consciousness seems to be useful for planning, manipulating images, self-modeling, etc. On the other hand, maybe this reflects the paucity of my human imagination in conceiving of ways to be intelligent without consciousness.

### 4.2 Sentient simulations

It may be that biological suffering is a drop in the bucket compared with digital suffering. The biosphere of a planet is less than Type I on the Kardashev scale; it uses a tiny sliver of all the energy of its star. Intelligent computations by a Type II civilization can be many orders of magnitude higher. So humans' sims could be even more troubling than their spreading of wild animals.

Of course, maybe there are ETs running sims of nature for science or amusement, or of minds in general to study biology, psychology, and sociology. If we encountered these ETs, maybe we could persuade them to be more humane.

I think it's likely that humans are more empathetic than the average civilization because

1. we seem much more empathetic than the average animal on Earth, probably in part due to parental impulses and in part due to trade, although presumably some of these factors would necessarily be true of any technologically advanced civilization
2. selection bias implies that we'll agree with our own society's morals more than those of a random other society because these are the values that we were raised with and that our biology impels us toward.

Based on these considerations, it seems plausible that there would be room for improvement through interaction with ETs. Indeed, we should in general expect it to be possible for any two civilizations or factions to achieve gains from compromise if they have diminishing marginal utility with respect to amount of control exerted. In addition, there may be cheap Pareto improvements to be had purely from increased intelligence and better understanding of important considerations.

That said, there are some downside risks. Post-humans themselves might create suffering simulations, and what's worse, the sims that post-humans run would be more likely to be sentient than those run by random ETs because post-humans would have a tendency to simulate things closer to themselves in mind-space. They might run nature sims for aesthetic appreciation, lab sims for science experiments, or pet sims for pets.

### 4.3 Suffering subroutines

Suffering subroutines may be a convergent outcome of any AI, whether human-inspired or not. They might also be run by aliens, and maybe humans could ask aliens to design them in more humane ways, but this seems speculative.

### 4.4 Black Swans

It seems plausible that suffering in the future will be dominated by something totally unexpected. This could be a new discovery in physics, neuroscience, or even philosophy more generally. Some make the argument that because we know so very little now, it's better for humans to stick around because of the "option value": If they later realize it's bad to spread, they can stop, but if they realize they should spread, they can proceed to reduce suffering in some novel way that we haven't anticipated.

Of course, the problem with the "option value" argument is that it assumes future humans do the right things, when in fact, based on examples of speculations we can imagine now, it seems future humans would probably do the wrong things much of the time. For instance, faced with a new discovery of obscene amounts of computing power somewhere, most humans would use it to run oodles more minds, some nontrivial fraction of which might suffer terribly. In general, most sources of immense power are double-edged swords that can create more happiness and more suffering, and the typical human impulse to promote life/consciousness rather than to remove them suggests that negative and negative-leaning utilitarians are on the losing side.

Still, waiting and learning more is plausibly Kaldor-Hicks efficient, and maybe there are ways it can be made Pareto efficient by granting additional concessions to suffering reducers as compensation.

## 5 What about paperclippers?

Above I was largely assuming a human-oriented civilization with values that we recognize. But what if, as seems mildly likely, Earth is taken over by a paperclip maximizer, i.e., an unconstrained automation or optimization process? Wouldn't that reduce suffering because it would eliminate wild ETs as the paperclipper spread throughout the galaxy, without causing any additional suffering?

Maybe, but if the paperclip maximizer is actually generally intelligent, then it won't stop at tiling the solar system with paperclips. It will want to do science, perform lab experiments on sentient creatures, possibly run suffering subroutines, and so forth. It will require lots of intelligent and potentially sentient robots to coordinate and maintain its paperclip factories, energy harvesters, and mining operations, as well as scientists and engineers to design them. And the paperclipping scenario would entail similar black swans as a human-inspired AI. Paperclippers would presumably be less intrinsically humane than a "friendly AI", so some might cause significantly more suffering than a friendly AI, though others might cause less, especially the "minimizing" paperclippers, e.g., cancer minimizers or death minimizers.

If the paperclipper is not generally intelligent, I have a hard time seeing how it could cause human extinction. In this case it would be like many other catastrophic risks – deadly and destabilizing, but not capable of wiping out the human race.

## 6 What if human colonization is more humane than ET colonization?

*If* we knew for certain that ETs would colonize our region of the universe if Earth-originating intelligence did not, then the question of whether humans should try to colonize space becomes less obvious. As noted above, it's plausible that humans are more compassionate than a random ET civilization would be. On the other hand, human-inspired computations might also entail more of what we consider to count as suffering because the mind architectures of the agents involved would be more familiar. And having more agents in competition for the light cone might lead to dangerous outcomes.

But for the sake of argument, suppose an Earth-originating colonization wave would be better than the expected colonization wave of an ET civilization that would colonize later if we didn't do so. In particular, suppose that if human values colonized space, they would cause only $-0.5$ units of suffering, compared with $-1$ units if random ETs colonized space. Then it would seem that as long as the probability $P$ of some other ETs coming later is bigger than 0.5, then it's better for humans to colonize and pre-empt the ETs from colonizing, since $-0.5 > -1 \cdot P$ for $P > 0.5$.

However, this analysis forgets that even if Earth-originating intelligence does colonize space, it's not at all guaranteed that human values will control how that colonization proceeds. Evolutionary forces might distort compassionate human values into something unrecognizable. Alternatively, a rogue AI might replace humans and optimize for arbitrary values throughout the cosmos. In these cases, humans' greater-than-average compassion doesn't make much difference, so suppose that the value of these colonization waves would be $-1$, just like for colonization by random ETs. Let the probability be $Q$ that these non-compassionate forces win control of Earth's colonization. Now the expected values are

$$-31 \cdot Q + -0.5 \cdot (1 - Q)$$

for Earth-originating colonization versus

$$-1 \cdot P$$

if Earth doesn't colonize and leaves open the possibility of later ET colonization.

For concreteness, say that $Q = 0.5$. (That seems plausibly too low to me, given how many times Earth has seen overhauls of hegemons in the past.) Then Earth-originating colonization is better if and only if

$$\begin{aligned} -1 \cdot 0.5 + -0.5 \cdot 0.5 &> -1 \cdot P \\ -0.75 &> -1 \cdot P \\ P &> 0.75. \end{aligned}$$

Given uncertainty about the Fermi paradox and Great Filter, it seems hard to maintain a probability greater than 75% that our future light cone would contain colonizing ETs if we don't ourselves colonize, although this section presents an interesting argument for thinking that the probability of future ETs is quite high.

What if rogue AIs result in a different magnitude of disvalue from arbitrary ETs? Let $H$ be the expected harm of colonization by a rogue AI. Assume ETs are as likely to develop rogue AIs as humans are. Then the disvalue of Earth-based colonization is

$$H \cdot Q + (-0.5) \cdot (1 - Q),$$

and the harm of ET colonization is

$$P \cdot (H \cdot Q + (-1) \cdot (1 - Q)).$$

Again taking $Q = 0.5$, then Earth-based colonization has better expected value if

$$\begin{aligned} H \cdot 0.5 + -0.5 \cdot 0.5 &> P \cdot (H \cdot 0.5 + -1 \cdot 0.5) \\ H - 0.5 &> P \cdot (H - 1) \\ P &> \frac{(H - 0.5)}{(H - 1)}, \end{aligned}$$

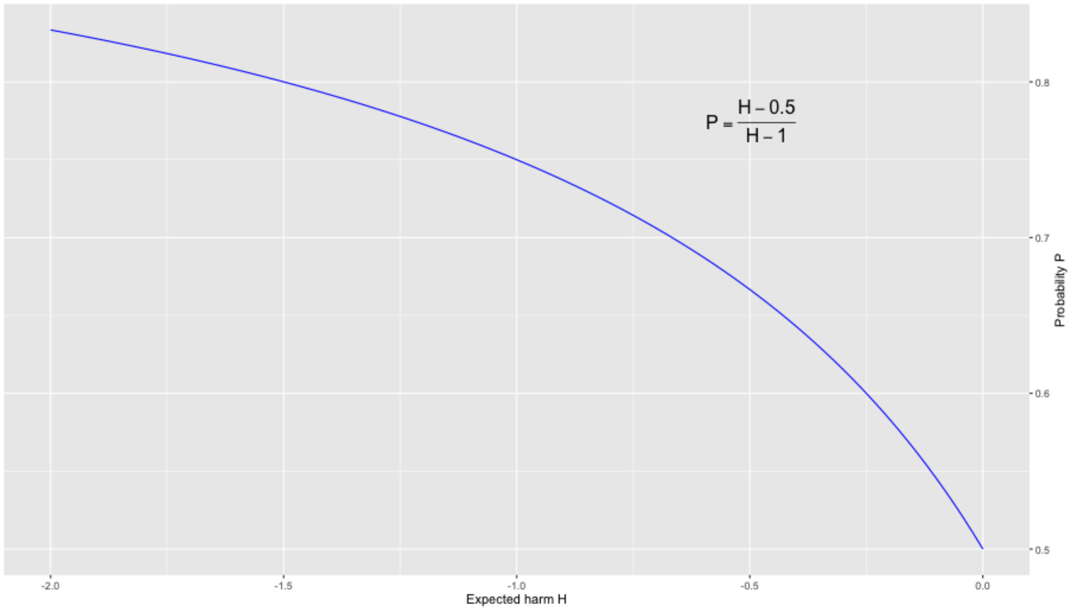

**Figure 1:** *Plot of threshold values for $P$ as a function of $H$*

where the inequality flips around when we divide by the negative number $(H - 1)$. **Figure 1** represents a plot of these threshold values for $P$ as a function of $H$.

Even if $H = 0$ and a rogue AI caused no suffering, it would still only be better for Earth-originating intelligence to colonize if $P > 0.5$, i.e., if the probability of ETs colonizing in its place was at least 50%.

These calculations involve many assumptions, and it could turn out that Earth-based colonization has higher expected value given certain parameter values. This is a main reason I maintain uncertainty as to the sign of Earth-based space colonization. However, this whole section was premised on human-inspired colonization being better than ET-inspired colonization, and the reverse might also be true, since computations of the future are more likely to be closer to what we most value and disvalue if humans do the colonizing.

## 7 Why we should remain cooperative

If technological development and space colonization seem poised to cause astronomical amounts of suffering, shouldn't we do our best to stop them? Well, it is worth having a discussion about the extent to which we as a society want these outcomes, but my guess is that someone will continue them, and this would be hard to curtail without extreme measures. Eventually, those who go on developing the technologies will hold most of the world's power. These people will, if only by selection effect, have strong desires to develop AI and colonize space.

Resistance might not be completely futile. There's some small chance that suffering reducers could influence society in such a way as to prevent space colonization. But it would be better for suffering reducers, rather than fighting technologists, to compromise with them: We'll let you spread into the cosmos if you give more weight to our concerns about future suffering. Rather than offering a very tiny chance of complete victory for suffering reducers, this cooperation approach offers a higher chance of getting an appreciable fraction of the total suffering reduction that we want. In addition, compromise means that suffering reducers can also win in the scenario ( 30% likely in my view) that technological development does prevent more suffering than it causes even apart from considerations of strategic compromise with other people.

Ideally these compromises would take the form of robust bargaining arrangements. Some examples are possible even in the short term, such as if suffering reducers and space-colonization advocates agree to cancel opposing funding in support of some commonly agreed-upon project instead.

The strategic question of where to invest resources to advance your values at any given time amounts to a prisoner's dilemma with other value systems, and because we repeatedly make choices about where to invest, what stances to adopt, and what policies to push for, these prisoner's dilemmas are iterated. In Robert Axelrod's tournaments on the iterated prisoner's dilemma, the best-performing strategies were always "nice," i.e., not the first to defect. Thus, suffering reducers should not be the first to defect against space colonizers. Of course, if it seems that space colonizers show no movement toward suffering reduction, then we should also be "provocable" to temporary defection until the other side does begin to recognize our concerns.

We who are nervous about space colonization stand a lot to gain from allying with its supporters – in terms of thinking about what scenarios might happen and how to shape the future in better directions. We also want to remain friends because this means pro-colonization people will take our ideas more seriously. Even if space colonization happens, there will remain many sub-questions on which suffering reducers want to have a say: e.g., not spreading wildlife, not creating suffering simulations/subroutines, etc.

We want to make sure suffering reducers don't become a despised group. For example, think about how eugenics is more taboo because of the Nazi atrocities than it would have been otherwise. Anti-technology people are sometimes smeared by association with the Unabomber. Animal supporters can be tarnished by the violent tactics of a few, or even by the antics of PETA. We need to be cautious about something similar happening for suffering reduction. Most people already care a lot about preventing suffering, and we don't want people to start saying, "Oh, you care about preventing harm to powerless creatures? What are you, one of those *suffering reducers*?" where "suffering reducers" has become such a bad name that it evokes automatic hatred.

So not only is cooperation with colonization supporters the more promising option, but it's arguably the only net-positive option for us. Taking a more confrontational stance risks hardening the opposition and turning people away from our message. Remember, preventing future suffering is something that everyone cares about, and we shouldn't erode that fact by being excessively antagonistic.

## 8 Possible upsides to an intelligent future

### 8.1 Black swans that don't cut both ways

Many speculative scenarios that would allow for vastly reducing suffering in the multiverse would also allow for vastly increasing it: When you can decrease the number of organisms that exist, you can also increase the number, and those who favor creating more happiness / life / complexity / etc. will tend to want to push for the increasing side.

However, there may be some black swans that really are one-sided, in the sense that more knowledge is most likely to result in a decrease of suffering. For example: We might discover that certain routine physical operations map onto our conceptions of suffering. People might be able to develop ways to re-engineer those physical processes to reduce the suffering they contain. If this could be done without a big sacrifice to happiness or other values, most people would be on board, assuming that present-day values have some share of representation in future decisions.

This may be a fairly big deal. I give nontrivial probability (maybe ~10%?) that I would, upon sufficient reflection, adopt a highly inclusive view of what counts as suffering, such that I would feel that significant portions of the whole multiverse contain suffering-dense physical processes. After all, the mechanics of suffering can be seen as really simple when you think about them a certain way, and as best I can tell, what makes animal suffering special are the bells and whistles that animal sentience involves over and above crude physics – things like complex learning, thinking, memory, etc. But why can't other physical objects in the multiverse be the bells and whistles that attend suffering by other physical processes? This is all very speculative, but what understandings of the multiverse our descendants would arrive at we can only begin to imagine right now.

### 8.2 Valuing reflection

If we care to some extent about moral reflection on our own values, rather than assuming that suffering reduction of a particular flavor is undoubtedly the best way to go, then we have more reason to sup-

[2] Of course, what's compelling to idealized-me would not necessarily be compelling to idealized-you. Value divergences may

port a technologically advanced future, at least if it's reflective.

In an idealized scenario like coherent extrapolated volition (CEV) (Yudkowsky, 2004), say, if suffering reduction was the most compelling moral view, others would see this fact.[2] Indeed, all the arguments any moral philosopher has made would be put on the table for consideration (plus many more that no philosopher has yet made), and people would have a chance to even experience extreme suffering, in a controlled way, in order to assess how bad it is compared with other things. Perhaps there would be analytic approaches for predicting what people would say about how bad torture was without actually torturing them to find out. And of course, we could read through humanity's historical record and all the writings on the Internet to learn more about what actual people have said about torture, although we'd need to correct for will-to-live bias and deficits of accuracy when remembering emotions in hindsight. But, importantly, in a CEV scenario, all of those qualifications can be taken into account by people much smarter than ourselves.

Of course, this rosy picture is *not* a likely future outcome. Historically, forces seize control because they best exert their power. It's quite plausible that someone will take over the future by disregarding the wishes of everyone else, rather than by combining and idealizing them. Or maybe concern for the powerless will just fall by the wayside, because it's not really adaptive for powerful agents to care about weak ones, unless there are strong, stable social pressures to do so. This suggests that improving prospects for a reflective, tolerant future may be an important undertaking. Rather than focusing on whether or not the future happens, I think it's more valuable for suffering reducers to focus on making the future better *if* it happens – by encouraging compromise, moral reflectiveness, philosophical wisdom, and altruism, all of which make everyone better off in expectation.

## Acknowledgments

A question by Rob Wiblin first inspired this piece. The discussion of cooperation with other value systems was encouraged by Carl Shulman. Initially I resisted his claim, but – as has often proved the case at least on factual and strategic questions – I eventually realized he was right and came around to his view.

## References

Bostrom, Nick. "The Future of Human Evolution." *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing.* Palo Alto, CA: Ria UP, 2004. 339-71. Web. 3 Mar. 2016 www.nickbostrom.com/fut/evolution.html.

Bostrom, Nick. "What Is a Singleton?" *Linguistic and Philosophical Investigations* 5.2 (2006): 48-54. Web. 3 Mar. 2016. www.nickbostrom.com/fut/singleton. html

Bostrom, Nick. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15.3 (2003): 308-3014. Web. 3 Mar. 2016. www.nickbostrom.com/astronomical/waste.html.

Pearce, David. "Extraterrestrial Life, the Fermi Paradox and the Hedonistic Imperative." *The Hedonistic Imperative.* N.p., n.d. Web. 3 Mar. 2016. www.hedweb.com/object32.htm

Thiel, Inari, Neil W. Bergmann, and William Grey. "A Case for Investigating the Ethics of Artificial Life?" *The University of New South Wales* 1 (2003): 276-87. Web. 3 Mar. 2016. http://espace.library.uq.edu.au/view/UQ:10754/A_Case_for_Inves.pdf

Tomasik, Brian. "Do Artificial Reinforcement-Learning Agents Matter Morally?" *ArXiv.org.* N.p., 30 Oct. 2014. Web. 03 Mar. 2016. http://arxiv.org/ abs/1410.8233v1.

Yudkowsky, Eliezer. *Coherent Extrapolated Volition.* San Francisco, CA: The Singularity Institute, 2004. Web. 3 Mar. 2016. http://intelligence.org/files/CEV.pdf.

remain after extrapolation. That said, there's no alternative better than compromising using a CEV-like approach, because if I try to defect and push my particular values, you'll just try to push yours, and we'll both be worse off in expectation.