Sequence summary

This sequence investigates the expected loss of value from non-extinction global catastrophes. This post is a criticism of the biases and ambiguities inherent in longtermist terminology (including ‘global catastrophes’). Part 2 lays out terms which I intend to use for the rest of this sequence, and which encourage less heuristic, more expected-value thinking. Part 3 lays out the structure of a proposed model which will inform the direction of my research for the next few months. If feedback on the structure is good, later parts will populate the model with some best-guess values, and present it in an editable form.

Introduction

Longtermist terminology has evolved haphazardly, so that much of it is misleading or noncomplementary. Michael Aird wrote a helpful post attempting to resolve inconsistencies in our usage, but that post’s necessity and its use of partially overlapping Venn diagrams - implying no formal relationships between the terms - itself highlights these problems. Moreover, during the evolution of longtermism, assumptions that originally started out as heuristics seem to have become locked in to the discussion via the terminology, biasing us towards those heuristics and away from expected value analyses.

In this post I discuss these concerns, but since I expect it to be relatively controversial and it isn’t really a prerequisite for the rest of the sequence so much as an explanation of why I’m not using standard terms, I would emphasise that this is strictly optional reading for the rest of the sequence, hence ‘Part 0. You should feel free to skip ahead if you disagree strongly or just aren’t particularly interested in a terminology discussion.

Concepts under the microscope

Existential catastrophe

Recreating Ord and Aird’s diagrams of the anatomy of an existential catastrophe here, we can see an ‘existential catastrophe’ has various possible modes:

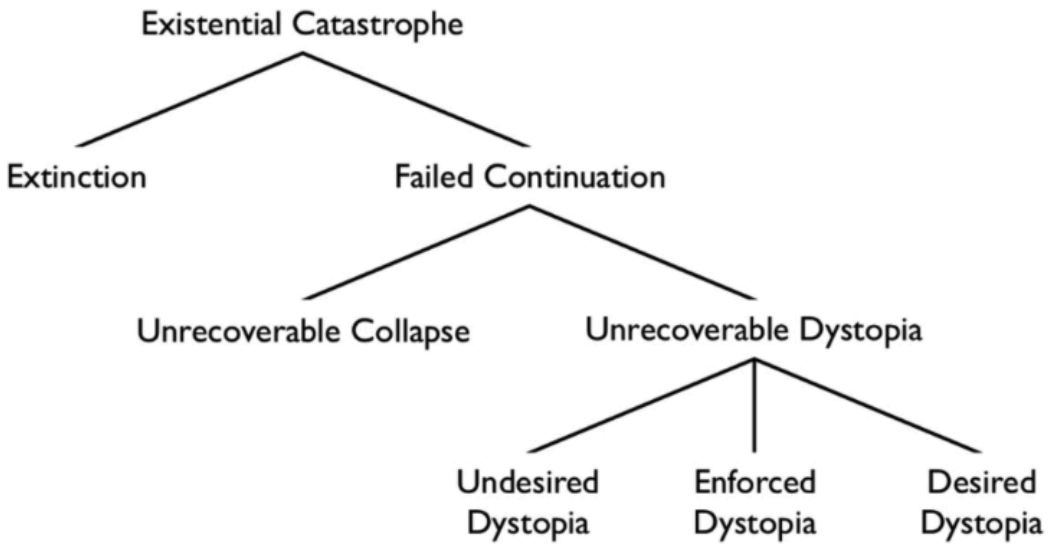
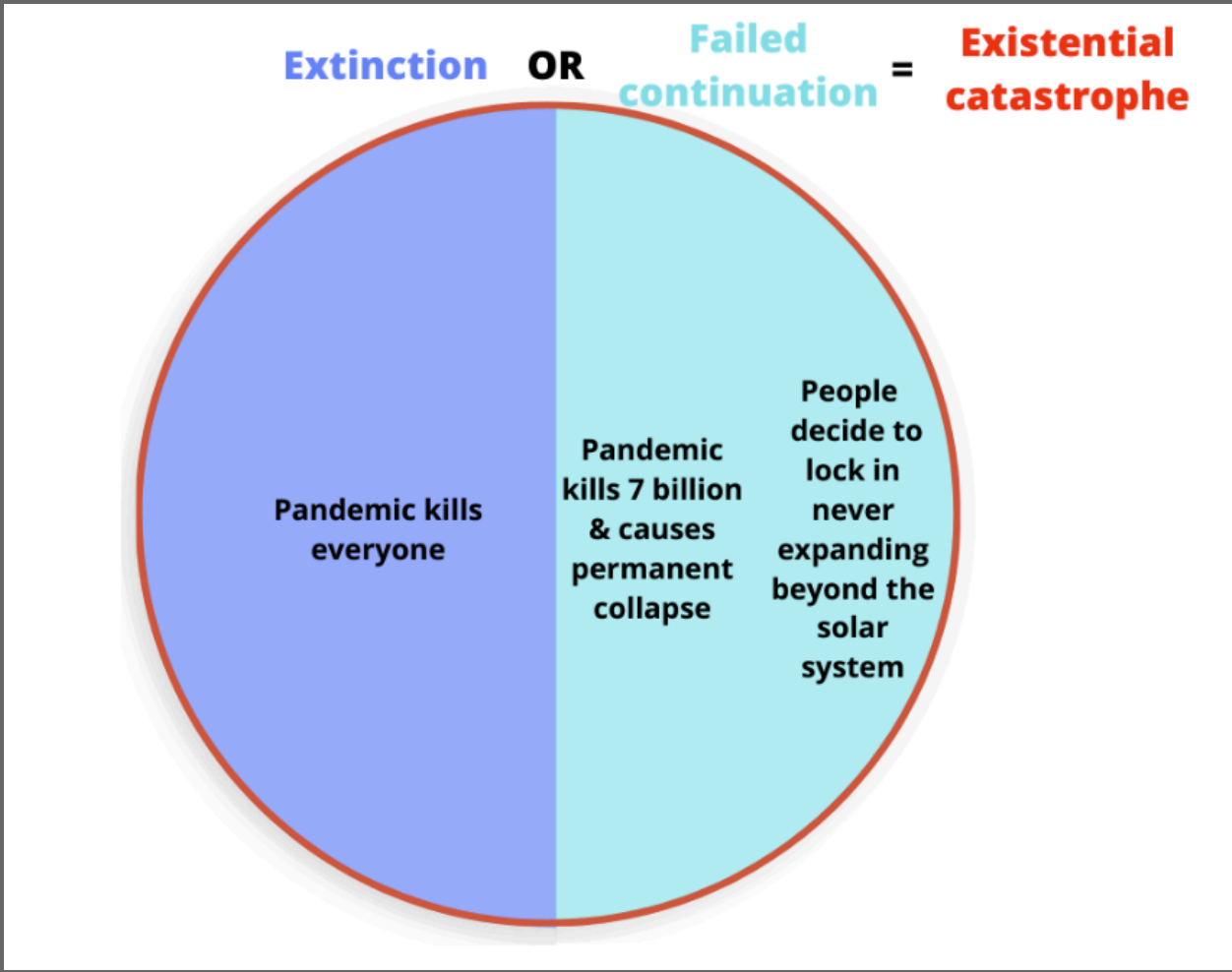


FIGURE 5.2 An extended classification of existential catastrophes by the kind of outcome that gets locked in.

Figure from The Precipice



Venn diagram figures all from Aird's post

It's the 'failed continuation' branch which I think needlessly muddies the waters.

An 'existential catastrophe' doesn't necessarily relate to existence…

In theory an existential catastrophe can describe a scenario in which civilisation lasts until the end of the universe, but has much less net welfare than we imagine it could have had.

This seems odd to consider an 'existential' risk - there are many ways in which we can imagine positive or negative changes to expected future quality of life (see for example Beckstead's idea of trajectory change). Classing low-value-but-interstellar outcomes as existential catastrophes seems unhelpful both since it introduces definitional ambiguity over how much net welfare must be lost for them to qualify, and since questions of expected future quality of life are very distinct from questions of future *quantity* of life, and so seem like they should be asked separately.

… nor involve a catastrophe that anyone alive recognises

The concept also encompasses a civilisation that lives happily on Earth until the sun dies, perhaps even finding a way to survive that, but never spreading out across the universe. This means that, for example, universal adoption of a non-totalising population ethic would be an existential catastrophe. I'm strongly in favour of totalising population ethics, but this seems needlessly biasing.

'Unrecoverable' or 'permanent' states are a superfluous concept

In the diagram above, Ord categorises 'unrecoverable dystopias' as a type of existential risk. He actually seems to consider them necessarily *im*permanent, but (in their existentially riskiest form) irrevocably harmful to our prospects, saying 'when they end (as they eventually must), we are much more likely than we were before to fall down to extinction or collapse than to rise up to fulfill our potential'[^xprecipice]. Bostrom imagines related scenarios in which 'it may not be possible to climb back up to present levels if natural conditions are less favorable than they were for our ancestors, for example if the most easily exploitable coal, oil, and mineral resources have been depleted.'

The common theme in these scenarios is that they lock humanity onto Earth, meaning we go extinct prematurely (as in, much sooner than we could have done if we'd expanded into the universe). Understood this way, the vast majority of the loss of value from either scenario comes from that premature extinction, not from the potentially lower quality of life until then or (following Bostrom's original calculation) from a delay of even 10 million years on the path to success. So at the big picture level to which an 'existential catastrophe' applies, we can class 'permanent' states as 'premature extinction'.

This doesn't hold for scenarios in which a totalitarian government rules over the universe until its heat death, but a) Ord's 'as they eventually must' suggests he doesn't consider that a plausible outcome, and b) inasmuch as it is a plausible outcome, it would be subject to the 'needn't relate to existence' criticism above.

Interpreting 'unrecoverable' probabilistically turns everything into a premature extinction risk

There seems little a priori reason to draw a *categorical* distinction between 'unrecoverable' and 'recoverable' states:

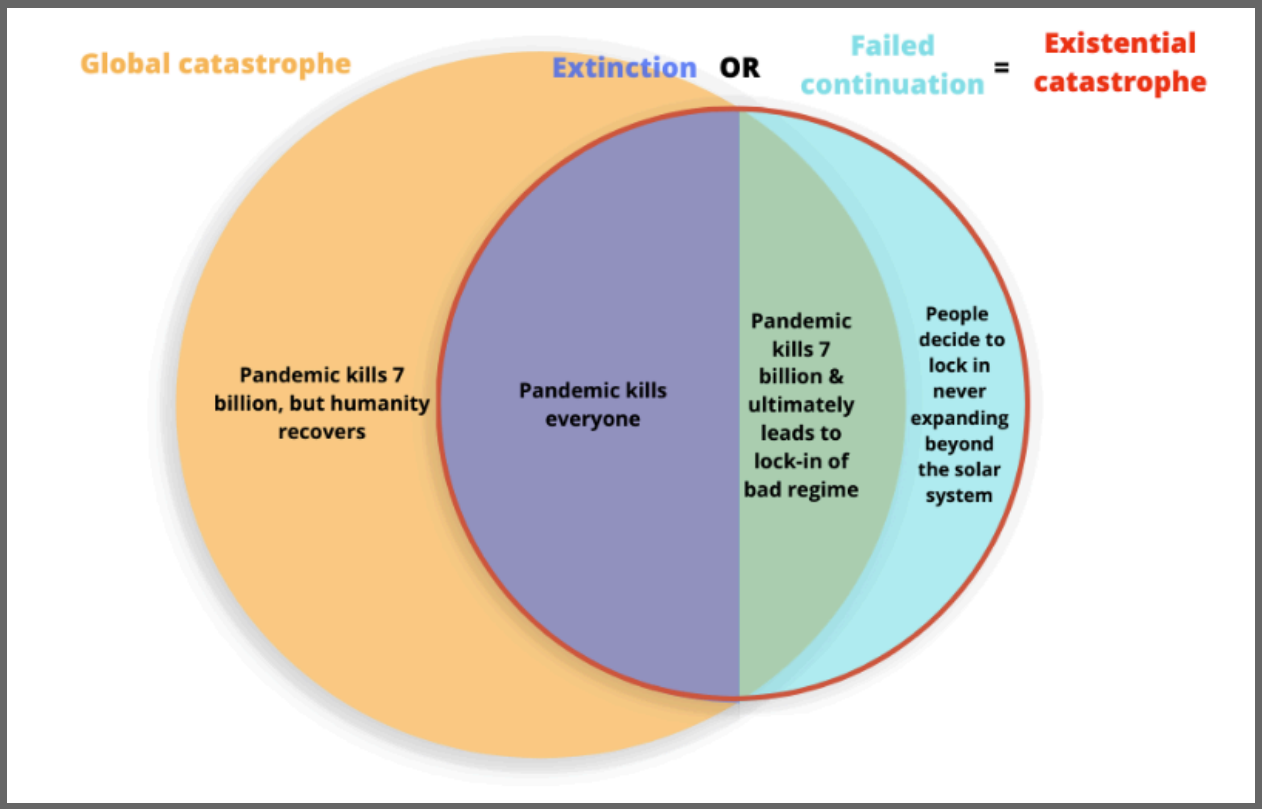
- Events that have widely differing extinction probabilities conditional on their occurrence might have widely differing probabilities of occurrence and/or widely differing tractability - so longtermists still need to do EV estimates to prioritise among them.
- It's extremely hard to estimate even our current premature extinction probability - on some assumptions, we're *already* in a state where it's high enough to give low expected value even given the possibility of astronomical value. So on any probabilistic definition

of an ‘unrecoverable’ state we might already be in one. Other people think a single serious setback would make success very unlikely.
- For categorical purposes it might not even be that valuable to estimate it - it’s hard to imagine a real-world resource-depletion or totalitarian scenario that felt so locked in that to a strict expected value maximiser the probability of success seemed low enough to give up on astronomical value.

This section isn’t meant to be a reductio ad absurdum - in Part 3 I’ll suggest a model on which even setbacks as ‘minor’ as the 2008 financial crisis carry some amount of premature extinction riskiness.

Global catastrophe

Aird’s essay, referencing Bostrom & Ćirković, considers a catastrophe causing 10,000 deaths or $10,000,000,000 of damage to be insufficient to qualify as ‘global’ and considers ‘a catastrophe that caused 10 million fatalities or 10 trillion dollars worth of economic loss’ to be sufficient - a definition shared by the UN. Aird thus presents the concept as an overlapping Venn diagram:



But these distinctions, as Bostrom & Ćirković observe, are fairly loose and not that practically relevant - and the concept of ‘recovery’ (see orange area) is importantly underspecified.

A global catastrophe isn't necessarily global…

Covid has caused about 6 million confirmed fatalities but probably over 20 million in practice *and* over 10 trillion dollars worth of economic loss, and I'm unsure whether in practice longtermists typically treat it as a global catastrophe.

Any absolute-value definition also misses something about relative magnitude - according to the World Bank, global GDP now is about 2.5 times its value 20 years ago, so 10 trillion dollars worth of economic loss around 2000 would have constituted 2.5 times the proportional harm it did today. And if the economy continues growing at a comparable rate, such a loss would barely register a century from now.

It's also vague about time: humanity has had multiple disasters that might have met the above definition, but most of them were spread over many years or decades.

… and focusing on it as a category encourages us to make premature assumptions

The longtermist focus on global catastrophic risks appears to be a Schelling point, perhaps a form of maxipok - the idea that we should maximise the probability of an 'ok outcome'. But though Bostrom originally presented maxipok as 'at best … a rule of thumb, a prima facie suggestion, rather than a principle of absolute validity', it has come to resemble the latter, without, to my knowledge, any rigorous defence of why it should now be so.

To estimate the counterfactual value of some event or class of events on the long-term future, we need to separately determine the magnitude of such events in some event-specific unit (here, fatalities or $ cost), and, separately, our credence in the long-term value of the event's various effects. Focusing on 'catastrophes' implies that we believe these two questions are strongly correlated - that we can be much more confident in the outcome of higher magnitude events. In the limit (extinction) this seems like a robust belief (but perhaps not a settled question), but it's less clear that it holds for lesser magnitudes. There might be good arguments to think the likelihood of long-term disvalue from an event that caused 10,000,000 deaths is higher than likelihood of long-term disvalue from one that caused 10,000, but the focus on global catastrophes bakes that belief into longtermist discussion before any argument has established such a strong correlation.

'Recovery' from catastrophe is a vague goal…

The concept of 'recovery' from a collapse is widely referred to in existential-risk-related discussion (see note 2.22 in last link) - but it's used to mean anything from 'recovering industry' through something like 'getting to technology equivalent to the modern day's'[^xrecover1], up to 'when human civilization recovers a space travel program that has the potential to colonize space'.

For any of these interpretations, the differential technological progress of a reboot would also make it hard to identify a technological state 'equivalent' to the modern world.[^xdartnell]

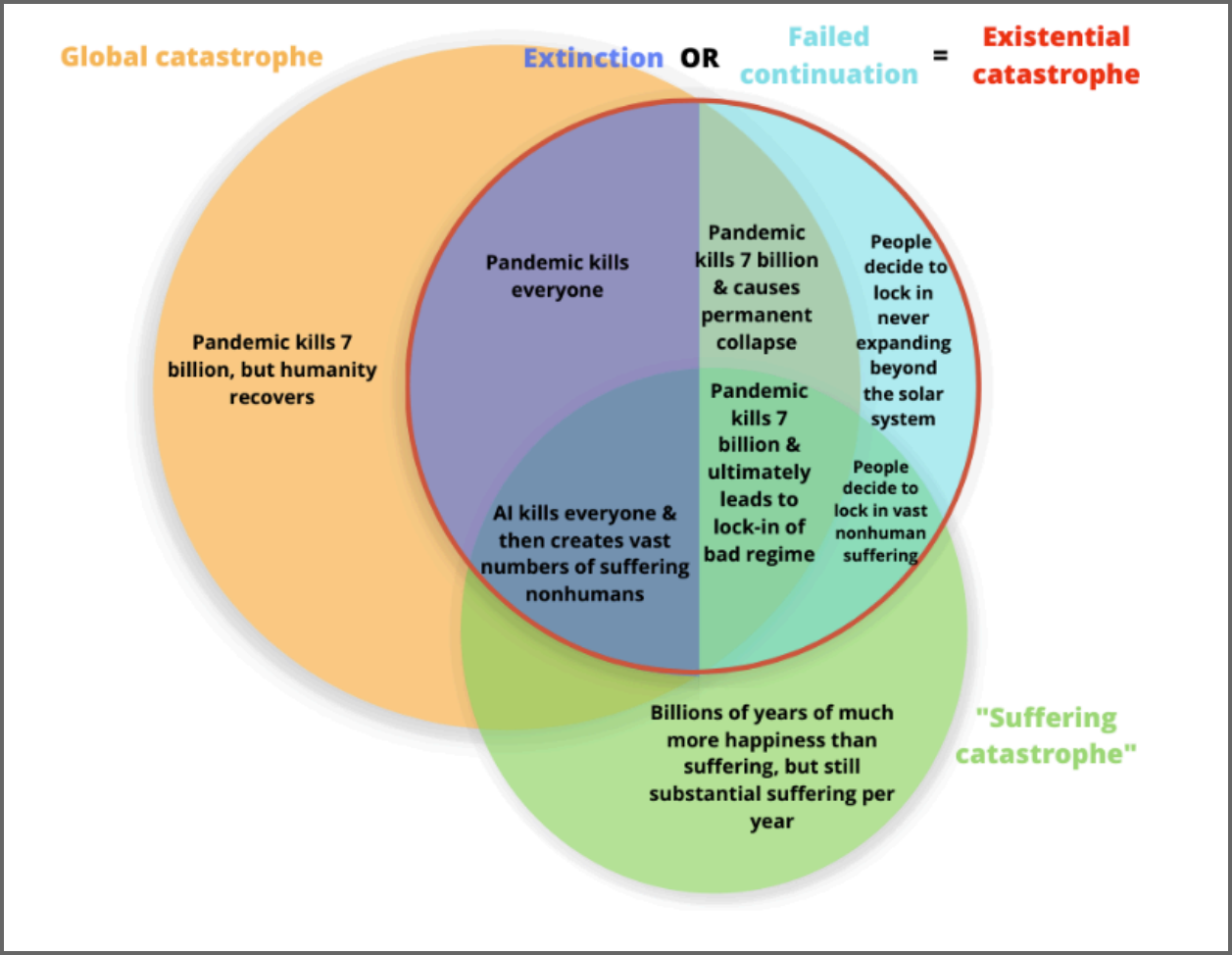
… and not the one we actually care about

As longtermists, we need to look at the endgame, which, per Bostrom's original discussion is approximately 'settling the Virgo Supercluster'[^xvalue]. The important question, given catastrophe, is not 'how likely is it we could reinvent the iPhone?', but 'how likely is it that civilisation would reach the stars?' This question, in addition to sounding cooler, captures two important concerns that as far as I've seen have only been discussed in highly abstract terms in collapse and resilience literature:

- the difference in difficulty for future civilisations of a) getting back to industrialisation, or wherever 'recovery' would leave us, and of b) getting from there to becoming robustly interstellar
- the possibility that those future civilisations might regress again, and that this could happen many times

Suffering catastrophe

Aird defines this as 'an event or process involving suffering on an astronomical scale'.



The definition of a suffering catastrophe depends on your moral values

For example a universe full of happy people who occasionally allow themselves a moment of melancholy for their ancestors would qualify under a sufficiently negative-leaning population ethic. Perhaps less controversially, a generally great universe with some serious long-term suffering would qualify under a less stringently negative ethic. So it seems often more helpful to talk about specific scenarios (such as systemic oppression of animals). And when we do talk about suffering catastrophes, or by extension S-risks, we should make explicit the contextual population ethic.

A mini-manifesto for useful expected-value terminology

To speak about a subject amenable to expected value analysis, I think as many as possible of the following qualities are useful, inspired by the discussion above:

- Having well defined formal relationships between the key concepts

- Using categorical distinctions only to reflect relatively discrete states. That is, between things that seem ‘qualitatively different’, such as plants and animals, rather than between things whose main difference is quantitative, such as mountains and hills
- Using language which is as close to intuitive/natural use as possible
- Using precise language which doesn’t evoke concepts that aren’t explicitly intended in the definition

Though there will usually be a tradeoff between the last two.

In the next post, I’ll give the terms that I intend to use for discussion around ‘catastrophes’ in the rest of the sequence, which adhere as closely as possible to these principles.

[^xrecover1] For an example of it meaning ‘recovering industry’, see Figure 3 in Long-Term Trajectories of Human Civilization. For an example of it meaning ‘getting to the modern day’, see the text immediately preceding and referring to that figure, which describes ‘recovering back towards the state of the current civilization’.

For another example of referring to the industrial revolution, see *What We Owe the Future*, which doesn’t define recovery explicitly, but whose section entitled ‘Would We Recover from Extreme Catastrophes?’ proceeds chronologically only as far as the first industrial revolution (the final mention of time in the section is ‘Once Britain industrialised, other European countries and Western offshoots like the United States quickly followed suit; it took less than two hundred years for most of the rest of the world to do the same. This suggests that the path to rapid industrialisation is generally attainable for agricultural societies once the knowledge is there.’)

In Luisa Rodriguez’s post specifically on the path to recovery, she seems to switch from ‘recovery of current levels of technology’ in her summary to ‘recovering industry’ in the discussion of the specifics.

[^xvalue] More specifically, settling it with a relatively benign culture. I tend toward optimism in assuming that *if* we settle it, it will be a pretty good net outcome, but, per a theme running throughout this post, that is a distinct question.

[^xprecipice] *The Precipice,* p225

[^xdartnell] In *The Knowledge*, Lewis Dartnell evocatively describes how ‘A rebooting civilization might therefore conceivably resemble a steampunk mishmash of incongruous technologies, with traditional-looking four-sail windmills or waterwheels harnessing the natural forces not to grind grain into flour or drive trip-hammers, but to generate electricity to feed into local power grids.’

Documents in this series:
Longtermist terminology has implicit biases
Name TBD
Modelling civilisation after a contraction