Concepts of existential catastrophe

Hilary Greaves (University of Oxford)

Global Priorities Institute | September 2023

GPI Working Paper No . 8-2023

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Concepts of existential catastrophe

Hilary Greaves

September 2023

Abstract

The notion of existential catastrophe is increasingly appealed to in discussion of risk management around emerging technologies, but it is not completely clear what this notion amounts to. Here, I provide an opinionated survey of the space of plausibly useful definitions of existential catastrophe. Inter alia, I discuss: whether to define existential catastrophe in *ex post* or *ex ante* terms, whether an *ex ante* definition should be in terms of loss of expected value or loss of potential, and what kind of probabilities should be involved in any appeal to expected value.

1. Introduction and motivations

Humanity today arguably faces various very significant *existential risks*, especially from new and anticipated technologies such as nuclear weapons, synthetic biology and advanced artificial intelligence (Rees 2003, Posner 2004, Bostrom 2014, Häggström 2016, Ord 2020). Furthermore, the scale of the corresponding possible catastrophes is such that anything we could do to reduce their probability by even a tiny amount could plausibly score very highly in terms of expected value (Bostrom 2013, Beckstead 2013, Greaves and MacAskill 2024). If so, then addressing these risks should plausibly be one of our top priorities.

An existential risk is a risk of an existential catastrophe. An existential catastrophe is a particular type of possible event. This much is relatively clear. But there is not complete clarity, or uniformity of terminology, over what exactly it is for a given possible event to count as an existential catastrophe. Unclarity is no friend of fruitful discussion. Because of the importance of the topic, it is worth clarifying this as much as we can. The present paper is intended as a contribution to this task.

The aim of the paper is to survey the space of plausibly useful definitions, drawing out the key choice points. I will also offer arguments for the superiority of one definition over another where I see such arguments, but such arguments will often be far from conclusive; the main aim here is to clarify the menu of options.

I will discuss four broad approaches to defining "existential catastrophe". The first approach (section 2) is to define existential catastrophe in terms of human extinction. A suitable notion of human extinction is indeed *one* concept that it is useful to work with. But it does not cover all the cases of interest. In thinking through the worst-case outcomes from technologies such as those listed above, analysts of existential risk are at least equally concerned about various other outcomes that do not involve extinction but would be similarly bad.

The other three approaches all seek to include these non-extinction types of existential catastrophe. The second approach appeals to loss of value, either *ex post* value (section 3) or expected value (section 4). There are several subtleties involved in making precise a definition based on expected value; I will suggest (though without watertight argument) that the best approach focuses on the consequences for expected value of "imaging" one's evidential probabilities on the possible event in question. The fourth approach appeals to a notion of the loss of humanity's potential (section 5). I

will suggest (again, without watertight argument) that when the notion of "potential" is optimally understood, this fourth approach is theoretically equivalent to the third.

The notion of existential catastrophe has a natural inverse: there could be events that are *as good* as existential catastrophes are bad. Ord and Cotton-Barratt (2015) suggest coining the term "existential eucatastrophe" for this inverse notion. Section 6 sets out the idea, and briefly discusses how useful we should expect this inverse notion to be in actual practice. Section 7 discusses the possibility of defining existential catastrophe in more purely descriptive terms. Section 8 summarises.

2. Defining "existential catastrophe" in terms of extinction

Among those who discuss existential risk, one concern is the possibility that humanity might go extinct in, say, the next century or two. This would plausibly be a massive-scale catastrophe not only from a partial point of view that has special concern for humans, but also by impartial lights, even if other species survived. The reason is that among species presently on Earth, humans are special. The specialness has two aspects. First, an especially high degree of sentience: humans have significantly higher capacity for well-being than individuals of most other species (Crisp 2003, p. 760; Kagan 2019, pp.42–52; McMahan 2002, p. 195–6; Schukraft 2020, p. 6–9; Vallentyne 2005, pp.405-6). Secondly, intelligence: because of the greater intelligence of humans, we alone have developed technologies facilitating massive rises in population size and in standards of living, and we alone have any remotely realistic prospect of settling parts of the universe beyond the Earth (Dyson 1999, chapter 3; Zubrin 1999; Smith & Davies 2012; Armstrong and Sandberg 2013; Kaku 2018; Baum et. al. 2019, section 7). It is plausible, therefore, that the extinction of humanity would lead to a much larger drop in overall expected value than the extinction of other species.[1]

As a first pass, we might try, therefore:

> ***Candidate Definition 1.1.*** *An event is an existential catastrophe iff it is the extinction of the human species.*

At least two things, however, are wrong with this definition.

First, it is inevitable that humanity will go extinct *at some point*, if only because of the eventual heat death of the universe. If humanity survived as long as was compatible with the facts of thermodynamics, its eventual extinction would not count as an existential catastrophe in any decision-relevant sense. What is catastrophic is not extinction *per se*, but rather *early* or *premature* extinction.

Second, we should not fixate on the disappearance of the human species *per se*. If *Homo sapiens* underwent continued evolution to such an extent that our successors came to count as members of a distinct biological species, that would not in and of itself be cause for concern. If *Homo sapiens* was replaced as the dominant species on Earth by some other type of entity, either of our own creation (genetically enhanced "posthumanity", artificial intelligence) or not (the result of mutations in a competitor species), that also need not be a catastrophe from the *impartial* point of view, provided

[1] This is not, of course, to belittle the project of broader species conservation. The statement that the extinction of humanity would entail a *much greater value loss than* the extinction of another species is a comparative one.

that the takeover species also possesses the morally relevant properties that made humans special in the first place (i.e., according to our above account of the "specialness", intelligence and sentience).[2]

This second consideration suggests replacing "human species", in our definition, with "intelligent and sentient life on Earth". However, it is also clear that "on Earth" is not quite what we want either. If Earth became uninhabitable but humans (or their relevantly similar successors) established comparably high-welfare settlements on other planets, the fact that no creatures of the specified type remained *on Earth* should not count as an existential catastrophe. Physical location has no intrinsic moral importance. On the other hand, if life on Earth went extinct but there continued to exist some unrelated alien civilisation quite elsewhere in the Universe, the latter fact should not prevent the extinction of Earth-life from counting as an existential catastrophe. We are interested in the absolute value loss involved in the events we contemplate, not the proportion of value lost in some grand cosmic sense.

The best definition of existential catastrophe *as an extinction event* therefore seems to be something like:

> ***Candidate Definition 1.2****. An event is an existential catastrophe iff it is the premature extinction of Earth-originating, intelligent, sentient life.*

As noted in section 1, however, this definition importantly fails to cover all cases of interest. Analysts of existential risk *are* concerned about possibilities of premature extinction, but they are *also* concerned about various other possibilities that seem similarly bad. The possibilities in question are generally cases in which the *size* and/or the *welfare* of the future population is massively reduced, in such a way that total welfare is massively reduced.[3] They include, for example, the possibilities that:

- All-out nuclear war (Sagan 1983, Robock et al 2007, Ellsberg 2017) or a global pandemic (Millet and Snyder-Beattie 2017) decimates the human population. Humanity survives, but with a massively reduced population size, and is reduced to subsistence conditions. Advanced technological civilisation never re-emerges.
- Advanced technology allows an oppressive totalitarian regime to take permanent control of the entire world, in such a way that most future people live in very low-welfare conditions (Caplan 2008)).
- Extreme climate change permanently and massively reduces the carrying capacity of the Earth, so that many fewer people can live at any given future time, without actually bringing forward the date of human extinction (Sherwood & Huber 2010).

Since the purpose of the project is to facilitate discussion also of this broader space of possible catastrophes, we need a definition that includes them in its scope.[4]

---

[2] A little more fundamentally, the desideratum is that there exist, throughout the long future, large numbers of high-welfare entities. One way this could come about is if a takeover species is both highly intelligent and highly sentient, and spreads itself throughout the universe (with high-welfare conditions). But other ways are also possible. For example, takeover by non-sentient AI together with extinction of humanity need not constitute an existential catastrophe if the non-sentient AI itself *goes on to create* large numbers of high-welfare entities of some other kind. (Thanks to Adam Bales for discussion of this point.)

[3] Why only 'generally'? Because it is at least in principle consistent with the notion of existential catastrophe that a sufficiently large loss of non-welfarist value could also count.

[4] Some might object to the use of the term "existential" in relation to this broader notion of catastrophe, on the grounds that the latter covers many cases that are not obviously about the *existence* of anything. I have some sympathy with this terminological thought, as far as it goes. For present purposes, though, the upshot is that ideally one would select a different term for the broader concept that I am here seeking to define, not that

A distinct (though related) point is that it is anyway controversial whether premature human extinction in fact *would* be enormously bad (or even whether it would be bad at all). This is for at least two reasons.

First, the claim plausibly hinges on controversial issues of axiology. One relevant issue here is population axiology (see, e.g., Greaves 2017 and references therein). The assumption that premature extinction is enormously bad follows quite naturally from a *totalist* population axiology, according to which the goodness of a state of affairs with respect to welfare is given by the *sum* of the lifetime welfare of all welfare subjects who ever live. But it is less clear what, for example, a "person-affecting" approach to population ethics says about the badness of extinction (Lenman 2002, Handfield 2018, Thomas 2022, Steele forthcoming).[5] Another relevant axiological issue is how to weigh suffering against happiness. Views that assign a very strong priority to reducing suffering over increasing happiness are less likely to underwrite the conclusion that premature human extinction would be bad (Gustafsson MS, Mogensen forthcoming).

Second, some hold the view that because of negative impacts of humans on members of other species, the early extinction of humanity would be net positive even by total utilitarian lights (May 2018; Benatar 2015).

Those who think premature extinction would not be enormously bad will not regard events meeting the conditions in Candidate Definition 1.2 as themselves very large-scale catastrophes. Yet even these people are likely to have a use for a notion of existential catastrophe in the more general sense, and are likely to regard at least some of the items on the above list (for example, technologically enabled totalitarian takeover) as qualifying.

If we assume that premature extinction *would* be enormously bad, then a natural approach to constructing a broader definition of existential catastrophe involves appealing to the badness of premature extinction as a benchmark. We can, that is, simply *define* existential catastrophe as any event that is close to as bad as premature human extinction (or worse), whether or not the event in question actually involves extinction. For simplicity, going forwards, I will make that assumption, and I will correspondingly pursue that approach. But the resulting definitions, of course, will succeed in picking out *very bad events* only if premature extinction indeed is enormously bad. Otherwise, existential catastrophe would need to be defined in a way that does not use that benchmark (probably: in a way that uses some different benchmark). I will leave open how this might best be done; the details seem likely to depend quite sensitively on what the reasons are for premature human extinction not being (very) bad.[6]

3. *Ex post* evaluative definitions

---

the broader concept requires no definition. An alternative reply is that the broader notion we seek *does* concern the existence of something: of a flourishing future of humanity, perhaps. In any case, here I stick with the term "existential" for continuity with existing discussions.

[5] This is of course not to say that premature extinction is enormously bad only if totalism is true. In particular, for discussions of the idea that premature human extinction might be enormously bad from quite different (non-welfarist) theoretical perspectives, see for example (Scheffler 2018, Riedener 2022).

[6] Just to give one example: If the reason for premature extinction not being bad relates to the negative impacts of humans on non-human animals, then a natural approach is to construct benchmark for the scale of catastrophe we intend to refer to by considering only the *intrinsic* value of human lives (and how much of that intrinsic value is lost in a premature extinction event).

In this section and the next, we consider definitions according to which, roughly, an existential catastrophe is any event that is about as bad as premature human extinction, or worse.

A significant choice point for this evaluative approach to defining a notion of catastrophe (existential or otherwise) is whether the "badness" that figures in the definition is *ex post*, or *ex ante* badness. In the present section, we explore what a definition in terms of *ex post* badness might look like; section 4 turns to definitions in terms of *ex ante* badness.

One might define existential catastrophe in terms of *ex post* badness as follows:

> ***Candidate Definition 3.1****: An existential catastrophe is an event which brings about the loss of a large fraction of the value of the future of humanity.*

As per the discussion in section 2, the "value of the future of humanity" is not intended here to exclude effects on on-human welfare subjects, but it *is* intended to exclude future value whose realisation has nothing to do with whether or not humanity continues (for example, pleasurable experiences had by uncontacted aliens). To capture this, we identify the "value of the future of humanity" with the amount of value that would be lost if humanity went immediately, but painlessly, extinct.[7]

If "existential catastrophe" is defined according to Candidate Definition 3.1, then we can in practice never be sure, *ex ante*, which possible events would be existential catastrophes. Suppose, for example, that an AI-facilitated global totalitarian takeover occurred, seeming highly likely to be permanent. It would be natural for an evaluator speaking shortly after such a takeover event to describe this as an existential catastrophe. But such an evaluator would not be in a position to know *with certainty* whether or not the takeover was permanent. Insofar as there remains a chance (however small) that the dictatorship turns out to be short-lived, on Candidate Definition 3.1, there remains that same chance that the takeover was not in fact an existential catastrophe.

Despite this, the notion of existential *risk* is of course an *ex ante* notion, even if "existential catastrophe" itself is defined in purely *ex post* terms. If there is a high chance that a given type of takeover event would bring about the loss of a large fraction of the value of the future of humanity, then, (even) on Candidate Definition 3.1, actions reducing the risk of such a takeover would in general count as reducing existential risk.

Candidate Definition 3.1 seems one reasonable way of defining existential catastrophe. In being a straightforwardly *ex post* definition, however, it is notably different from extant definitions of existential catastrophe in the (admittedly rather small) literature.

There might be good reasons for this. In particular, one might worry that Candidate Definition 3.1 leads to false positives. The reason is that even our most trivial actions have some chance of bringing about arbitrarily terrible long-run future effects: for example, by affecting the identities of future people (Lenman 2000:347, Greaves 2016:314). Let E be a particular event of my helping somebody to cross the road. Suppose that E occurs, and that via identity-affecting effects, this brings about a permanent totalitarian takeover a century later. On the face of it, E counts as an existential catastrophe according to Candidate Definition 3.1. But it seems at least arguably a little strange to call it a catastrophe - let alone an existential one - that I helped somebody to cross the road, even if things do turn out in this especially unfortunate way. It is perhaps more natural to describe E as an

[7] Because the "value of the future of humanity" is thus defined as a value *difference*, there is no supposition of a privileged zero point on the value scale involved in talking about *fractions* of that value.

intrinsically benign event that, unfortunately, brought about some *other* event that was an existential catastrophe (via an utterly unpredictable route).

While "arguably a little strange", that E (in the above example) counts as an existential catastrophe might however be simply the right result; nothing problematic obviously follows from it. For example, the fact (if it is a fact) that E is an existential catastrophe does not imply that it would be rational for us to spend enormous resources preventing E from occurring. *Ex ante*, the causal connections between E and its negative far-future consequences are utterly inaccessible to us, and it is equally possible that not-E is an existential catastrophe (for similar, and similarly unpredictable, reasons; cf. Greaves 2016, section 4)).

One might think that events like E are disqualified from counting as existential catastrophes on the grounds that there is no (or: no significant) loss of value intrinsic to the event itself. In this way of thinking, only goings-on that are intrinsic to a given event are relevant for an event's status as a catastrophe (existential or otherwise); call this the "intrinsicality requirement" (IR). According to IR, the Boxing Day Tsunami counts as a catastrophe (rather than merely as the *cause* of a catastrophe) partly because the deaths that it caused are naturally counted as part of the event that we pick out when we say "Boxing Day Tsunami" in the first place. On this line of thinking, returning to our above example of helping somebody to cross the road, there we should count E as *causing* an existential catastrophe, but not as itself *being* one.

In the context of more ordinary catastrophes (such as the Boxing Day Tsunami), IR is relatively plausible.[8] The problem with IR specifically in the context of *existential* catastrophes, however, is that it is in fact characteristic of the latter quite generally that most of the value loss they bring about is spread across the whole of the future of humanity (or, in the case of catastrophes of premature extinction: the whole of the time period that would *otherwise* have corresponded to the future of humanity). When we classify early extinction of humanity, population collapse or technologically enabled totalitarian takeover as an existential catastrophe, we are not primarily thinking of the disvalue that is directly involved in that event itself, on any remotely natural construal of the boundaries of the event (although that disvalue is of course also large by any normal standards). Rather, we are primarily thinking already of the negative effects of the event in question across the whole of the future. Applying IR as a constraint on a definition of existential catastrophe, therefore, would disqualify the paradigm examples from falling under the resulting concept.

If we do not want to count E as an existential catastrophe but we also reject IR, a natural route is to construe (or amend) the "brings about" clause in the definition in such a way that the highly unpredictable, chaotic causal chain that leads from E to an enormous loss of value does not count as "bringing about" in the relevant sense. We might, for example, require that the causal chain in question be reasonably *robust* to small variations in initial conditions, or (relatedly) *predictable*.

If we make the latter move, however, then we are well on the way to an *ex ante* definition. It may be more natural then simply to define existential catastrophe in terms of loss of expected value, rather

[8] I say only "relatively plausible". Even in these cases, IR might not in the end survive scrutiny. A global pandemic such as COVID-19 involves both bad effects that are plausibly intrinsic to the pandemic itself (such as the deaths that occur during and as part of the pandemic), and potentially also serious negative after-effects (such as educational, economic and cultural consequences playing out over several years). It seems at least arguably proper that the extent and magnitude of the latter form part of the assessment of what scale of catastrophe (if any) the pandemic in question should be counted as.

than in terms of a "reasonably predictable" loss of (*ex post*) value. This is the route we explore in section 4.

4. Defining "existential catastrophe" in terms of loss of expected value

For this next approach, the idea is to replace "value" in Candidate Definition 3.1 with some notion of *ex ante* value. Different approaches to decision theory correspond to different notions of *ex ante* value. For concreteness, taking the lead from orthodox decision theory, I will employ specifically the notion of *expected* value. But nothing will hinge on the details of the expected value formula; those who favour some alternative decision theory can simply replace "expected value" throughout with their favoured notion of *ex ante* value.

4.1 The basic definition

Adjusting Candidate Definition 3.1 accordingly gives us[9]:

> ***Candidate Definition 4.1.*** *An existential catastrophe is an event which brings about the loss of a large fraction of the expected value of the future of humanity.*

In the end, something very close to this might be the best definition of existential catastrophe. But at least two further clarifications are required.[10]

The first centres on the term "expected". Expected value is a matter of probability-weighted average value *with respect to some particular probability distribution*. If talking in terms of expected value (or any other *ex ante* evaluation), therefore, one key question is what determines which is the relevant probability distribution. For example, a straightforward subjectivist approach would appeal to the evaluator's own subjective credences (whatever they might be). But one can also consider more objective criteria for picking out the relevant probability distribution.

The second source of unclarity is the term "brings about". Since the value function on possible worlds is fixed, to bring about a change in expected value, an event must bring about a change in the probability distribution. But there are several, importantly distinct, senses in which this could be the case. In one sense, an event E "brings about" a given shift in the relevant probability distribution if a causal consequence of that event's occurring is that the shift in question occurs over time (as a temporal change in, for example, the evaluator's subjective or evidential credences or the objective chances). On an alternative approach, the shift is instead a matter of *conditioning or imaging* the existing probability distribution in question on E (i.e., on the proposition that the event in question occurs).[11]

These two issues are related. In particular, how one resolves the first affects what is a plausible way of resolving the second. Without attempting to exhaust the space of possibilities, we will consider two approaches that seem at least minimally plausible. On the first such approach (section 4.2), we take the relevant probability distribution to be a matter of evidential probability; we will argue that

---

[9] A similar definition is discussed by Ord and Cotton-Barratt (2015, p.2). Ord and Cotton-Barratt write simply "a large fraction of expected value". The refinement "of the future of humanity", however, is required by the considerations discussed in section 3.

[10] Nothing in the issues discussed in the remainder of this section is specific to the notion of *existential* catastrophe; the discussion would apply equally to definitions of more common-or-garden levels of catastrophe in terms of expected value loss.

[11] We will say more below about the nature of the conditioning and imaging operations. Both are operations that map a pair consisting of a probability distribution p and an event E to a probability distribution p_E that assigns probability 1 to E; but the two operations are distinct, in ways that are potentially important for our purposes.

in that case, "brings about" must be understood in terms of imaging (or perhaps conditioning). On the second approach (section 4.3), we take the relevant probability distribution to be an objective chance distribution, and understand "brings about" in terms of a change over time in those objective chances.

4.2 Evidential probabilities

Suppose first that we take the relevant probability distribution to be the evaluator's evidential probabilities.[12]

The *evidential probability* of a proposition p, relative to evidence E, is the objective likelihood of p on evidence E; we are then especially interested in the case in which E is the total evidence a given evaluator has (at a given time).[13] The picture is that while it might be difficult for limited creatures like us to determine which probabilities these are, nonetheless there is a fact of the matter.[14] Because the body of evidence possessed by the evaluator at a time changes over time, so also does the relevant probability distribution. It is also, of course, relative to the identity of the evaluator: your evidence might be relevantly different from my evidence. For each time t and evaluator S, we have a corresponding body of evidence $E_{S,t}$, and hence an evidential probability distribution $p_{S,t} = p(.|E_{S,t})$.[15]

If these are the kind of probabilities we are dealing in, we must understand "brings about" in terms of an abstract operation, rather than in terms of temporal transition. This is because coupling evidential probabilities to the latter notion of "bringing about" results in both false negatives and false positives, vis-à-vis the question of which events count as existential catastrophes. False negatives: An event of premature human extinction is a paradigm example of an existential catastrophe, but the causal consequence of such an event is not to bring about a negative shift in the evaluator's probability function (rather, the evaluator herself is annihilated). For a second example of a false negative: takeover by a sufficiently dire totalitarian dictatorship that is *in fact* permanently locked-in should count as an existential catastrophe, even if (for lack of visible *evidence* of the lock-in) humanity retains a perpetual rational belief that recovery is just around the corner. False positives: If a totalitarian dictatorship takes over at t_1 but humanity only *gains evidence* that the dictatorship is locked in at t_2, the event of one's gaining the evidence is not itself an existential

[12] A close cousin of this approach instead appeals to the evaluator's *actual subjective credences*, whether or not they are rational in light of the available evidence. We will discuss evidential probabilities for definiteness and ease of exposition, but most of the discussion in section 3.4 applies in the same way on this subjectivist approach.

[13] One usually does not go far wrong (at least) by thinking of the evidential probabilities, relative to a given body of evidence E, are the credences that it would be rational for an evaluator whose total evidence was E to have. For discussion of whether this identification is precisely correct, see e.g. Williamson 2000, Eder forthcoming.

[14] According to the *uniqueness hypothesis*, for any given body of evidence precisely *one* credal state is rationally permitted (Kopec and Titelbaum 2016). Discussion of evidential probabilities is simpler if the uniqueness hypothesis is assumed, but the notion of evidential probability does not essentially presuppose uniqueness. In the absence of uniqueness, some indeterminacy may arise, if a given proposition that implicitly or explicitly invokes evidential probability (for example, the proposition that unaligned AI takeover would be an existential catastrophe) is true relative to some rationally permitted credal states but false relative to others. Similar remarks apply to the issue of whether it is rationally permitted and/or required to have precise credences (Elga 2010, White 2010, Schoenfield 2012).

[15] We will simply accept that ascriptions of "existential catastrophe" status are evaluator-relative, without attempting to pin down who the evaluator is. One might, for example, take the relevant evidential probabilities to be the individual speaker's. One could also, perhaps, consider the evidential probabilities applicable to "humanity as a whole" at a given time.

catastrophe (rather, it is evidence that an existential catastrophe has already occurred at the earlier time).[16]

What we want is a probability-shifting operation that takes a probability distribution and a proposition P as input, and outputs a (generally distinct) probability distribution that assigns probability 1 to P, *even if P could occur without anyone being aware of, or being in a position to be aware of, its occurrence*. In sections 4.2.1 and 4.2.2, we consider two candidates for this probability-shift operation.

4.2.1 Conditioning

One might take the relevant probability-shifting operation to be conditionalization. In that case, our definition of existential catastrophe would be[17]:

> ***Candidate definition 4.2.*** *An event C is an existential catastrophe, relative to time t and evaluator S, iff starting from probabilities $p_{S,t}$, conditioning on C reduces expected value by a large fraction of the previously expected value of the future of humanity. That is, iff*
>
> $$E[V, p_{S,t}] - E[V, p_{S,t}(.|C)] > \alpha(E[V, p_{S,t}] - E[V, p_{S,t}(.|X)]),$$
>
> *where X is the proposition that humanity goes immediately but painlessly extinct, and α<1 is a large fraction.*[18]

A consequence of using conditioning as the probability-shift operation, however, is that any event that would be strong *evidence* that an existential catastrophe will occur (relative to the evaluator and time in question) itself tends to count as an existential catastrophe (relative to that evaluator and time). A special case of this is that if C is an event of *the evaluator's gaining* strong evidence that there will be an existential catastrophe, then C itself will often count as an existential catastrophe.

Consider, for example:

> ***Secret lock-in.*** *A totalitarian dictatorship takes over the world. Life consequently becomes grim for the vast majority of humans. A resistance movement is organised, and humanity retains hope that the dictatorship will be successfully overthrown in the lifetimes of current people. In fact, however, the dictatorship is enabled by powerful AI technology, and consequently is permanently locked in. This lock-in is only known to those in the dictatorship's inner core. It is possible that, the dictators aside, humanity will live out the rest of its days still*

[16] The phenomenon of false negatives would loom still larger if we used subjective rather than evidential probabilities, as we can see by considering non-rational processes of belief revision. While a bad day or a tingle in the little finger might cause an agent to become massively more pessimistic about the prospects for the future of humanity, said bad day (or finger-tingle) should not thereby count as an existential catastrophe.

[17] Here we assume that "the expected value of the future of humanity" is itself to be defined in terms of the results of *conditioning* on the proposition X that humanity goes (immediately and painlessly) extinct, rather than in terms of the results of *imaging* on X. This seems natural, and perhaps required by a robust notion of consistency, *if* (as here) one is using conditioning (rather than imaging) with respect to the event C whose status is being evaluated. It is, however, not compulsory. It is also arguably misguided, for the same reasons that it is arguably misguided to appeal to conditioning on C rather than imaging on C (section 3.4).

[18] The value of α representing "a large fraction" is, of course, somewhat vague or arbitrary. This is desirable, since the concept of existential catastrophe is itself vague or arbitrary in the same way. One might take α to be 90%, but 50% would also be reasonable.

rationally clinging to its objectively forlorn hope of recovery. Let D be the possible event of humanity's discovering that this lock-in has occurred.

At least intuitively, it seems that in *Secret Lock-in*, the existential catastrophe itself occurs when the dictatorship takes over, and that D itself would merely *inform* humanity at large that a catastrophe had occurred. D, according to this intuition, would not *itself be* an existential catastrophe. If this intuition is correct, this is a problem for Candidate Definition 4.2, since the latter implies that D itself *is* an existential catastrophe.

This ("intuitive") verdict could be reasonably resisted. As with the case of the objectively catastrophic road-crossing discussed in section 3, it does not seem that counting D as an existential catastrophe would suggest any inappropriate resource allocation. In this case, that is because there is plausibly nothing we could do to decrease the likelihood of D without also decreasing the likelihood that we obtain *good* news about the dictatorship's prospects for being permanent.[19]

My own sense, for whatever it is worth, is that it would however be preferable if we could say that D is evidence of a catastrophe without itself being a catastrophe. And in fact, there is a straightforward way of securing that verdict, as we next explain.

4.2.2 Counterfactual imaging

The problem (if there is a problem) is that in *Secret Lock-in*, the event D is *evidence* of a catastrophe, without *causally bringing about* that things are terribly bad.

We are in familiar territory. In the classic Newcomb's problem (Nozick 1969; Peterson 2017, ch. 9), the proposition that one will one-box is *evidence* of good outcomes without *causally bringing about* good outcomes. We obtain the (arguably unwanted) result that one-boxing is rationally superior to two-boxing if we evaluate actions by expected value relative to probabilities that are *conditioned on* the proposition that the action in question is performed. We obtain the (arguably wanted) result that two-boxing is rationally superior to one-boxing if we instead evaluate actions by means of expected value relative to probabilities that are *imaged* on the proposition that the action in question is performed, for a suitable notion of imaging. Here, "imaging" is an operation that, like conditioning, maps a pair consisting of a probability distribution and a proposition P to a probability distribution that assigns probability 1 to P; but that differs from conditioning on important other details. Specifically, to obtain the (arguably) desired verdict on Newcomb's problem, we want an operation of counterfactual imaging: when imaging on A, prior credence in each non-A world is distributed between A-worlds w according to the (non-backtracking) chance that w would obtain if A happened.[20]

For an arbitrary probability function p and proposition A, we write p^A for the probability function that results from counterfactual imaging p on A. Returning to the case in hand, we might then try:

Candidate definition 4.3. *An event C is an existential catastrophe, relative to time t and agent A, iff starting from probabilities $p_{A,t}$, counterfactual imaging on C reduces expected value by a large fraction of the previously expected value of the future of humanity. That is, iff:*

$$E[V, p_{A,t}] - E[V, p^C{}_{A,t}] > \alpha(E[V, p_{A,t}] - E[V, p^X{}_{A,t}]),$$

[19] For some discussion of the possibility or otherwise of "intentionally biassed inquiry", i.e. inquiry deliberately designed to asymmetrically yield possibilities of supporting a preferred conclusion, see Salow (2018).

[20] Imaging is described in more detail by Lewis (1981, esp. sections 6 and 9).
On the distinction between backtracking and non-backtracking counterfactuals, see Lewis (1979).

where X is the proposition that humanity goes immediately but painlessly extinct, and $\alpha<1$ is a large fraction.

Candidate Definition 4.3 gives the intuitively desirable result in *Secret Lock-in*. Let p be the evidential probabilities corresponding to humanity's evidence before the lock-in is discovered. p, then, assigns probability 1 to the proposition that a totalitarian takeover has occurred, but assigns low probability to the proposition that the resulting dictatorship is locked in. We now consider the consequences of imaging p on the proposition D that humanity discovers compelling evidence of lock-in. We are thus led to consider counterfactuals of the form: if D, then the prospects for the future of humanity would be thus-and-so.

It is tempting, *but crucially misleading*, to think: if D, then probably the dictatorship *is* locked in, in which case the prospects for the future of humanity are bleak. Such a counterfactual has a ring of strong plausibility, *but it is a backtracking counterfactual*. For backtracking counterfactuals, the key observation is instead that since gaining evidence that lock-in has occurred does not *causally bring about* lock-in, counterfactual imaging on D does not boost the probability assigned to lock-in: it merely shifts probability mass from non-D worlds with (resp., without) lockin to D-worlds with (resp., without) lockin. (The result will be a probability distribution that places significant mass on worlds in which there is *misleading* evidence of lock-in.)

A potentially problematic issue with any view that defines existential catastrophe in terms of evidential probabilities concerns the status of events that are intuitively existential catastrophes, but whose occurrence is *already part of the evidence* at the time of evaluation. Suppose, for example, that a permanent totalitarian lock-in L occurred in 2010 and that this has been known since 2010. Suppose we are evaluating in 2023 whether or not L was an existential catastrophe. On Candidate Definitions 4.3 and 4.4, *relative to 2023* L is *not* an existential catastrophe. That is because the evidential probabilities relative to 2023 already reflect the information that L occurred, and so, starting from *those* probabilities, neither conditioning on L nor imaging on L occasions any change in expected value. There seems therefore to be a danger that "L was an existential catastrophe", uttered in 2023, would be false. If so, this result seems bad.

A direct response to this issue would supply an explanation of how and why the danger just referred to is merely apparent. For example, one might hold that whether or not a given event counts as an existential catastrophe is a time-relative affair. On this account, what is false in 2023 is that L *is* an existential catastrophe. But, the account continues, if the question is whether L *was* an existential catastrophe – say, at the time it occurred – then the relevant probabilities are the evidential probabilities that prevailed *before L occurred* (or was determined to occur). And there is no obstacle to L's counting as an existential catastrophe relative to the latter.

Besides this direct type of response, the more general point is that insofar as there is a problem here, it is anyway of a piece with the "problem of old evidence" for a Bayesian view of scientific confirmation (Glymour 1980:145-50; Earman 1992, chapter 5). This suggests a "partners in guilt" response: there surely must be some adequate response to the problem of old evidence, and there is every reason to think that an analogous response to the above "problem of old catastrophes" will be equally adequate, even without settling the details of what the response is.

4.3 Macrotheoretical chances

A quite different way of avoiding the (arguably unwanted) implication that the discovery event D (in *Secret lock-in*) is an existential catastrophe is to take a particular type of *objective chances*, rather than evidential probabilities, to be the relevant probability distribution.

The rough idea is as follows. Recall our observation that *if* the relevant probability distribution is based on a notion of *credence* (whether rational credence, or subjective credence), *then* one wants to define existential catastrophe in terms of conditioning or imaging the existing probabilities on the proposition that the candidate catastrophe event occurs, rather than in terms of the shift in probability distribution that is brought about in real time by that event. This was in order to be able to conclude, for example, that an event of permanently locked in totalitarian dictatorship is still an existential catastrophe even if nobody has any evidence (or any credence) that the dictatorship has been locked in. However, if one instead takes the relevant probabilities to be objective chances, there is no such obstacle to defining existential catastrophe in terms of (actual or counterfactual) shifts in the relevant probability distribution over time. The objective chance of a dictatorship being permanent can be close to 1 even if no-one has evidence of (or credence in) its permanence.

The suggestion would then be:

> ***Candidate definition 4.4.*** *Let p_t be the objective chance function at time t. An existential catastrophe occurs between t_1 and t_2>t_1 iff expected value with respect to p_{t2} is less than expected value with respect to p_{t1} by a large fraction of the expected value of the future of humanity as of t_1: that is, iff*
>
> $$E[V, p_{t_1}] - E[V, p_{t_2}] > \alpha(E[V, p_{t_1}] - E\left[V, p^X{}_{t_2}\right]),$$
>
> *where $p^X{}_{t_1}$ is the probability function obtained by imaging p_{t_1} on the proposition X that humanity goes immediately but painlessly extinct, and α<1 is a large fraction.*

The question for this type of definition is which notion of objective chance it uses. Candidate Definition 4.4 is unpromising if "objective chance" is read as *microphysical* chance — that is, the probabilities that figure in the true theory of microphysics, given the true microphysical state at the time in question. (For example, though not only, because if microphysics is deterministic then the probability of the actual future is at all times 1.)

On one view, the chances of the *most fundamental* description of physical reality are the only chances there are (Schaffer 2007). This view, however, seems mistaken: besides microphysics, we also have a number of higher-level theories that involve probabilities with every claim to be bona fide chances, and that are uncontroversially indeterministic (Glynn 2010). Folk macrophysics, for example, fruitfully recognises a sense in which (what we usually call) a fair coin would still have a chance of one half of landing Heads on (what we usually call) a fair coin-flip, even if microphysics were deterministic. Similarly in spirit, our best theories of the behaviour of human and political systems (such as they are) might recognise the sense in which the chance of all-out nuclear war was "somewhere between one out of three and even" (according to then President John F Kennedy) at the time of the Cuban Missile Crisis, again, even if microphysics is deterministic.

We might, then, define existential catastrophe with respect to the chances that appear in some significantly higher level of theory than microphysics – the theory (or constellation of theories) that we operate with when doing "ordinary macroscopic description of the world", perhaps. In addition to recognising the sense in which the Cuban Missile Crisis plausibly involved non-trivial existential risk, such an account also seems able to give an adequate treatment of the other cases we have discussed. In *Secret lock-in*, for instance, the true macrotheoretical description of the world would include details of the dictatorship's possession of advanced AI technology and the implications of this, whether or not anyone was in a position to know the truth of that description. We would then get the (arguably desirable) results that the advent of the lockin is an existential catastrophe, while

humanity's discovering said lockin is not (since such a discovery would not depress the *objective chance* of recovery any lower than it already was).

A potential difficulty with a "macrotheoretical chance" approach lies in specifying more precisely which "macrotheory" is the one relative to which existential catastrophe is defined. There are, of course, many theories that are "significantly higher-level than microphysics" – including, for example, chemistry, biology, folk physics, psychology, and various theories of social science. Which is the relevant level of description for present purposes, and in virtue of what is that level of description privileged?

In light of this challenge, there seem to be three ways things could pan out. First, perhaps the question just posed has an adequate direct answer: perhaps, that is, one level of description is somehow privileged for present purposes. Second, perhaps there is no *privileged* answer, but this does not doom the approach: there are many levels of description that might be of interest, and the notion of existential catastrophe is itself simply relative to level of description. Each of these possibilities seems prima facie plausible, though either would require further development. Third, one could of course abandon the appeal to macrotheoretical chances, perhaps returning to a definition in terms of evidential probabilities as discussed above.

It seems to me somewhat unclear whether a definition in terms of evidential probabilities (and imaging), or instead one in terms of macrotheoretical chances (and shifts therein over time), is preferable to the other. My own inclination is to prefer the former, on the grounds that the issue of level-relativity seems to render the latter unhelpfully complicated. But readers who are less averse to such complication, and/or more worried by the "problem of old catastrophes", might reasonably prefer a definition in terms of macrotheoretical chances.

5. Defining "existential catastrophe" in terms of loss of potential

The majority of the current literature on existential risk takes neither of the routes above, but instead defines existential catastrophe in terms of *loss of potential*. For example: [21]

> ***Candidate Definition 5.1 (Bostrom).***[22] *An existential catastrophe is the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development. (Bostrom 2013)*
>
> ***Candidate Definition 5.2 (Ord).*** *An existential catastrophe is the destruction of humanity's longterm potential. (Ord 2020, p.37)*

The key choice point for a definition of this type is what the "potential" of humanity, or of Earth-originating intelligent life, should be taken to be. There is considerable flexibility here, making "potential"-based definitions a broad church. Here, I will consider three ways to make a potential-

[21] The two definitions quoted here are essentially equivalent. The most-cited definition of existential catastrophe is probably Bostrom's. Ord omits the explicit reference to extinction on grounds of superfluousness – if humanity goes extinct, then its potential is lost. The fact that Bostrom writes "Earth-originating intelligent life" while Ord writes "humanity" superficially looks like a substantive difference, but Ord makes clear that he intends "humanity" as a shorthand for something like this more accurate but more cumbersome phrase (2020, p.38).

[22] Strictly speaking, Bostrom does not directly give a definition of existential catastrophe: what he writes is that an "An existential *risk* is one that *threatens* the premature extinction…" (*ibid*., emphasis added). However, I take it that the above mild reconstruction is faithful to his intention.

based definition of existential catastrophe more precise. Of these, I will suggest, two are undesirable, while the third renders a potential-based definition equivalent to one in terms of expected value.

First, one might identify "potential" with the value of the best future that remains possible (Ord 2020, p.301, n.4). This quantity varies in time insofar as the set of futures that are possible-as-of-t varies with the time t. Presumably, a future that was impossible at an earlier time cannot become possible at a later time. But one might think the reverse can happen. At the personal level, for example, if you elect not to go to a party, then the future in which you enjoy a lifelong relationship with the soulmate you could have met there perhaps becomes forever inaccessible. If so, then although potential (in this sense) cannot increase, it might decrease. In the best-case scenario, potential remains constant. An existential catastrophe would then be an event bringing about a drastic, and by definition irreversible, reduction in the value of the best-case scenario.

This first way of making "potential" more precise is inadequate for our purposes. One problem is that in reality, there might not be any cases in which a previously possible future becomes literally *impossible*. Whatever happens, if nothing else, a quantum fluctuation restoring things to the alternative state of affairs is always *possible*. (For example, even if humanity goes extinct, it is still *possible* for it to be restored, down to the last atom, by such a fluctuation. More prosaically, even if you eschew that party, it is still *possible* (if unlikely) that you meet and bond with the person in question via some other route.[23]) So, on this first definition, humanity's potential (if it is even well-defined) would not change in time even in the face of the types of events that we seek to classify as existential catastrophes.

A second problem is that even if (*contra* the worry explored in the previous paragraph) there are cases in which potential in this first sense is reduced, these seem unlikely to cover all the cases of interest. Consider, for example:[24]

> ***Dictatorship with narrow escape hatch***. A totalitarian global dictatorship takes over the world. Structural features and access to advanced technology ensure that the regime is highly likely to last until the extinction of humanity. However, there remains a small chance (around 1%) that the regime might be successfully overthrown.

The presence of the "escape hatch" in this example is a positive feature, but it seems insufficient to disqualify *Dictatorship with narrow escape hatch* from counting as an existential catastrophe.[25] It is, however, enough to ensure that "potential" *in the sense we are currently considering* has not drastically reduced. I conclude that a definition of existential catastrophe based on the notion of potential should not understand "potential" in this first way.

The general lesson of *Dictatorship with narrow escape hatch* is that events that *shift large amounts of probability mass* from massively better to massively worse futures must sometimes, and then for that reason, count as existential catastrophes. We are now into territory that is familiar from more general discussions of decision-making under uncertainty. Following the lead of orthodox decision

---

[23] In response to such fanciful suggestions as quantum fluctuations, it may be tempting to draw a distinction between (something like) 'bare' possibilities and 'live' possibilities, where the latter but not the former are (something like) realistically plausible scenarios. But in order to be relevant for present purposes, such a distinction must be a matter of the magnitude of the probabilities involved. This naturally leads to the alternative ways of making "potential" precise – most naturally, as discussed below, in terms of expected value.

[24] A similar example is discussed by Ord and Cotton-Barratt (2015, pp.2-3) and alluded to by Ord (2020, p.40).

[25] On the *ex post* definition discussed in section 3, one would say that the takeover event in Dictatorship with narrow escape hatch is *probably* an existential catastrophe (with probability around 99%).

theory, if we wish to take account of the significance of shifting probabilities, it is natural think in terms of the *ex ante* value (for example, the expected value) corresponding to a given probability distribution, rather than simply in terms of a binary property of possibility/impossibility attaching to futures. This is the second way we might make "potential" precise. But this, of course, would return us to definitions of the type considered in Section 4.

A third approach is a variant on the second.[26] We might be especially concerned with how well things would be expected to go *if humanity behaves appropriately*. On this way of thinking, humanity's potential is not a matter of expected value *simpliciter*, but might be identified with its expected value *conditional on* the proposition that humanity follows the best possible course of action.

A definition along these lines is considered by Ord (2020, p.301, n.4). This seems to me, however, not a fruitful route. A significant number of "existential catastrophe" scenarios that generate concern are catastrophes precisely because they are scenarios in which we *do not expect* humanity to follow (anything like) the best possible policy. Scenarios of permanently locked-in global totalitarian dictatorship, for example, are of this type. The effect of conditioning on humanity's following the best possible policy is simply to exclude these scenarios that in fact we wish to include.

My own conclusion from this section is that while there is nothing in principle wrong with definitions of existential catastrophe that appeal to the notion of "potential", when the term "potential" is understood as it then should be, such definitions are theoretically equivalent to ones that directly use the language of expected value.

6. Non-evaluative definitions

The point of discussing existential catastrophes is essentially evaluative: these possible events are important to discuss because they would be so *bad*. It does not immediately follow, however, that the definition of "existential catastrophe" itself has to include evaluative concepts.

Let us say that a definition is *purely descriptive* if it uses no evaluative concepts. In this paper, we have seen one type of definition that is purely descriptive, viz. definitions in terms of premature extinction (discussed in section 2). Definitions in terms of loss of value (section 3) or expected value (section 4) are obviously not purely descriptive (indeed, it is tempting to characterise them as *purely evaluative*, though I will not attempt to make the latter notion precise). Whether a definition in terms of potential (section 5) is purely descriptive depends on how "potential" is cashed out, though, as we have seen, the more plausible ways of cashing it out make this definition at least partly evaluative too.

Owen Cotton-Barratt and Toby Ord have both suggested (in correspondence) that there might be some advantages to a purely descriptive definition. One reason one might think this is that in some discussion contexts (in particular, in some scientific contexts), explicit use of evaluative concepts is discouraged; a purely descriptive definition of "existential catastrophe" might then make it easier for existential risk to receive discussion in these contexts. A second reason stems from the fact that purely descriptive vs. evaluative definitions have different properties in the presence of evaluative disagreement.

[26] This approach is not essentially tied to the definition of "potential"; it could equally be regarded as simply a further possible refinement of a definition in terms of expected value.

It is not clear to me that either of these considerations in the end favours trying to construct a purely descriptive definition of "existential catastrophe".[27] It is also worth noting that, apart from definitions in terms of premature extinction (which, I complained, are too narrow to cover all the cases of interest), no extant definition of existential catastrophe has this property.

There is, however, an important general point in this vicinity that might ground a preference for a definition with more descriptive content than (for example) Candidate Definition 4.1. In general, given a partly evaluative definition, one can construct a definition of a nearby concept by replacing or supplementing the existing evaluative language with descriptive language that one believes picks out the important actual cases falling under the evaluative definition. In doing so, one makes use of one's own evaluative and empirical beliefs: another person, with different such beliefs, might think there is more distance between the more purely evaluative and the more descriptive definitions in question.

"Going descriptive" in this way will often *de facto* involve narrowing a definition. Rather than trying to include all possible sources of disvalue, we "zoom in", picking out, in more descriptive terms, those that seem to be the most important in the actual situation. This can be a helpful tool for focussing attention on the important cases.

In the context of existential risk, one example concerns the notion of permanence. The definitions discussed in section 3-5 do not, as they stand, require that the loss of expected value or potential involved in an existential catastrophe be permanent.[28] To see that we might want to impose such a condition, consider the following (unrealistic) case:

> ***Short-lived astronomical suffering*.** A powerful sadist creates a truly astronomical number of people, in the 21st century, all with lives of unmitigated torture. The number of these people is greater than the expected total number of people from the year 2200 onwards. The people created by the sadists have no descendants, the sadist dies, and after 2200 things return to normal.

The point of *Short-lived astronomical suffering* is that if the number of created sufferers is large enough, this scenario could correspond to just as large a reduction in expected value as premature human extinction. It would therefore count as an existential catastrophe according to a definition of the type considered in section 4. However, it might be desirable to *exclude* cases like this from our definition. This is because the cases of real-world concern do not in fact include any in which the large reduction of expected value comes about via an enormous but short-lived "negative spike" in the graph of expected well-being against time (as noted above, the case is *unrealistic*). The paradigm cases all rather concern cases in which expected welfare is depressed "permanently", i.e., across the whole of the future of humanity's (possible) existence. If we seek to exclude cases like *Short-lived astronomical suffering* from the definition, so as to focus attention more narrowly on the cases that

---

[27] Regarding scientific discussion: The natural move seems to be to allow the notion of existential catastrophe *itself* to be evaluative, and for the associated evaluative discussion to motivate conducting separate (and less evaluatively loaded) analyses of those possible events that seem most likely to instantiate that evaluative concept (for example, lock-in scenarios involving artificial intelligence).
Regarding evaluative disagreement: either parties with different evaluative views will agree on which events would count as existential catastrophes but disagree on how bad (if at all) an existential catastrophe would be, or such parties will disagree on the first matter while agreeing on the second. Neither of these options seems clearly superior to the other, in terms of facilitating useful discussion across important evaluative disagreement.

[28] Bostrom's definition (our Candidate Definition 5.1), of course, already includes a condition of permanence. While Ord's definition (our Candidate Definition 5.2) does not include permanence in its official wording, Ord makes clear in his discussion that he does also intend to impose such a condition.

are of real concern, we might thereby want to restrict our definition by including a suitable condition of permanence.[29]

One might also seek to go further in this direction (for example, also building into a definition a suitable condition of 'irreversibility', or a condition concerning the source of the catastrophe in e.g. inadequately managed new technology). In all such cases, the tradeoff one faces is between informativeness on the one hand, and simplicity and breadth of applicability on the other.

7. Existential eucatastrophe

Existential catastrophes are things that are enormously *worse* than the status quo. A natural question is whether there is also a useful structurally similar notion with the opposite valence. Ord and Cotton-Barratt suggest the term "existential eucatastrophe" for such a thing.

This possibility is naturally suggested, in particular, by definitions of existential catastrophe in terms of expected value.[30] By modifying definition 4.1, for example, we naturally obtain:

> ***Existential eucatastrophe (definition).*** *An existential eucatastrophe is an event which brings about a large increase in expected value, comparable to or greater than the existing expected value of the future of humanity.*

As Ord and Cotton-Barratt note, historical candidates for existential eucatastrophes include any event of humanity's passing a "great filter" – an event that was *ex ante* unlikely, but that was a prerequisite for the development of intelligent life. Past examples plausibly include the genesis of life, of multicellular life and of intelligence (Carter and McCrea 1983; Snyder-Beattie et. al. 2021), and the industrial revolution (Pinker 2019; Mokyr 2017: 339–41; Allen 2009: 272–5). Future examples might include the successful settlement of space (Bostrom 2003), human enhancement (Bostrom and Savulescu 2009) and "friendly" advanced artificial intelligence (Muehlhauser and Bostrom 2014: 43).

As noted in passing above, having a notion of existential eucatastrophe in our toolkit potentially helps with giving a complete account of *Dictatorship with narrow escape hatch*. In the event that the unlikely escape materialises, we might describe the resulting scenario as one of existential catastrophe followed by a corresponding existential eucatastrophe. We thereby capture, in one natural way, the sense in which the initial onset of dictatorship was something worth trying to avoid at almost all costs by the lights of the evidence available at the time, while also capturing the fact that *as things subsequently turned out* its effects were nowhere near as bad as expected.

---

[29] On one reading, *Short-lived astronomical suffering* itself involves a "permanent reduction in the expected value of the future of humanity". For, even as things stand after 2200, if the expected value we are calculating is still the expected value of the future of humanity *from the 2020s onwards*, then it will remain the case after 2200, and forever, that the sadist's actions caused the loss of a large fraction of that expected value. However, there is another sense in which the loss is not permanent: if we consider the expected value of the future of humanity from t onwards, and (for the purposes of considering permanence) ask whether a large fraction of the expected value of *that* is lost relative to every time t, then *Short-lived astronomical suffering* does not count as a permanent loss.

[30] It is also possible, though perhaps a bit more awkward, to formulate a definition of existential eucatastrophe in terms of the notion of potential. One could consider, for example, the notion of "an event which brings about a drastic increase in humanity's longterm potential, comparable to or greater than the magnitude of humanity's existing longterm potential."
In either case, in place of "comparable to or greater than" the existing value or potential, we could instead require the increase to be "many times" the existing value or potential. It is not immediately clear which way of defining eucatastrophe (if either) would be more fruitful. Thanks to Toby Ord for this observation.

Just as it is extraordinarily valuable to reduce the chances of existential catastrophes, so it may also be extraordinarily valuable to increase the chances of existential eucatastrophes. Whether this is so in practice depends on how tractable the project of "increasing the chance of existential eucatastrophe" is – whether there are any such eucatastrophes that are at all plausible, and if so, how much we can do (and at what cost) to increase the chances that they occur.

Ord and Cotton-Barratt implicitly suggest that many existing projects count as attempts to increase the chances of existential eucatastrophe:

> Some people are trying to identify and avert specific threats to our future – reducing existential risk. Others are trying to steer us towards a world where we are robustly well-prepared to face whatever obstacles come – they are seeking to increase existential hope [i.e., the chances of existential eucatastrophes]. (*ibid.*, p.4)

The suggested inherent connection between "robust preparations" and eucatastrophe, however, is obscure: such preparations seem to be in principle *neutral* between a focus on existential catastrophe vs. eucatastrophe. In particular, insofar as "what is to come" is additional dangers of existential catastrophe, then such preparations would amount to (indirect, general) existential risk reduction efforts.

A simple abstract model might help to fix ideas. Let S be some proposition S such that E(S) >> E(not S) >> E(X). Suppose that the current probability of S is neither very close to 0, nor very close to 1. Current expected value, then, lies between E(S) and E(not S), and is not very close to either. In general, it will then tend to be the case that conditioning or imaging on S increases expected value by an appreciable fraction of the expected value of humanity, *and also* that conditioning or imaging on not-S *decreases* expected value by an appreciable fraction of the expected value of humanity. In that case, S counts as an existential eucatastrophe *and* not-S counts as an existential catastrophe; so action that brings about an increase in the probability of S counts *simultaneously* as decreasing existential risk, and increasing existential hope.

Two conditions will, in some cases, defeat the equivalence that this seems to suggest between decreasing existential risk and increasing existential hope.

First, if the probability of (say) S is already very close to 1, then conditioning on S would not bring about much increase in expected value, while conditioning on not-S would bring about a significant decrease in expected value. In that case, not-S would count as an existential catastrophe, but S would not count as an existential eucatastrophe; S is, in such a case, close to being simply the status quo. The situation is of course symmetric: if the probability of not-S is already very close to 1, then S would count as an existential eucatastrophe, but not-S would not count as an existential catastrophe.

Second, it is natural (though not obligatory) to work in this context with a fairly thick notion of "event", according to which not just any set of possible worlds counts as an "event". It is natural, for example, to regard a sufficiently specific space-settlement achievement (for example, the establishment of the first self-sufficient settlement on a planet other than Earth) as an event, while not regarding the proposition that humanity never achieves any such settlement as an event (the latter is, rather, a *non-event* in the literal sense). One might on these grounds regard space settlement as an existential eucatastrophe, while not regarding failure to ever achieve space settlement as an existential catastrophe. (Though there could, even with this more robust notion of event, be an *event that makes it very significantly less likely that space settlement will ever be achieved*; the latter could still count as an existential catastrophe.)

The question of to what extent there are in fact plausible eucatastrophes that we can effectively intervene on seems to me somewhat open, and I will not attempt significantly to advance discussion of that here.

8. Summary and conclusions

In this paper, I have considered various possible ways in which one might define "existential catastrophe". *One* useful notion to focus on is that of the premature extinction of "humanity", broadly construed. For this to be a useful focus, "humanity" must indeed be construed broadly, so as to avoid fetishization of species boundaries and so as to secure (insofar as is desired) impartiality between humans and other relevantly similar entities. Generally, where "humanity" appears in a definition of existential catastrophe, it is to be read as an abbreviation for "Earth-originating intelligent sentient life".

However, scenarios of premature extinction are far from the only locus of concern. Many of the same new and emerging technologies that give rise to concerns about premature human extinction also give rise to similarly serious concerns about similarly bad non-extinction scenarios. It is important to have a concept that also includes these other scenarios.

Analysts of existential risk have offered various definitions that are intended to capture the idea of "either premature extinction, or another event that is relevantly similar and (in particular) similarly bad or worse". Some of these appeal directly to the notion of expected value loss, while the others appeal to loss of "potential". I have argued that when "potential" in this context is optimally understood, it is equivalent to expected value. It seems to me therefore less confusing to simply define existential catastrophe in terms of expected value directly, eschewing talk of "potential" as far as the official definition goes.

On an expected value approach, the *rough* idea is to define existential catastrophe as an event that brings about a large loss of expected value. There are then various choice points regarding how to make this rough idea more precise. One concerns which are the relevant probabilities (subjective probabilities, evidential probabilities, or objective chances), for the purpose of determining *expected* value. A second and related choice point concerns whether to understand "brings about" temporally, or by means of conditioning or imaging on the proposition that the event in question occurs.

Within the "expected value" approach, my own preferred definition of existential catastrophe is:

> ***An existential catastrophe is an event that brings about the permanent loss of a large fraction of the expected value of humanity,***

where:

- "Fraction" is understood relative to a baseline scenario in which humanity goes immediately but painlessly extinct.
- The probabilities in question are evidential probabilities.
- "Brings about" is understood in terms of counterfactual imaging those probabilities on the proposition that the event in question occurs.

An alternative approach omits "expected" in the definition above, defining existential catastrophe in *ex post* rather than *ex ante* terms.

There is significant room for reasonable difference on one's preferred definition of existential catastrophe. The main desideratum is that we attain clarity on which definition is being used, to

whatever extent clarity is useful for the discussion at hand. The main aim of the present paper has been to assist in attaining such clarity.

References

Allen, R. (2009) *The British industrial revolution in global perspective*. Cambridge: Cambridge University Press.

Armstrong, S. and Sandberg, A. (2013) "Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox." *Acta Astronautica* 89. doi.org/10.1016/j.actaastro.2013.04.002.

Baum, S., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., Maas, M., Miller, J., Salmela, M., Sandberg, A., Sotala, K., Torres, P., Turchin, A. and Yampolskiy, R. (2019) "Long-term trajectories of human civilization." *Foresight* 21 (1): 53–83. doi:10.1108/FS-04-2018-0037

Beckstead, N. (2013) *On the overwhelming importance of shaping the far future*. Ph.D. dissertation, Rutgers, The State University of New Jersey.

Benatar, D. (2015) "The misanthropic argument". In *Debating procreation: Is it wrong to reproduce?* New York: Oxford University Press. p. 78–121.

Bostrom, N. (2003) "Astronomical waste: The opportunity cost of delayed technological development." *Utilitas*, 15(3), 308-314.

Bostrom, N. (2014) *Superintelligence: Paths, dangers, strategies.* Oxford: Oxford University Press.

Bostrom, N. (2013) "Existential risk prevention as global priority." *Global Policy* 4.1: 15-31.

Bostrom, N. and Savulescu, J. (eds.) (2009) *Human enhancement*. Oxford: Oxford University Press.

Caplan, B. (2011) "The totalitarian threat." In Nick Bostrom & Milan Cirkovic (eds.), *Global catastrophic risks* (pp. 507–519). Oxford: Oxford University Press.

Carter, B., & McCrea, W. H. (1983) "The anthropic principle and its implications for biological evolution [and discussion]." *Philosophical Transactions of the Royal Society of London, Series A: Mathematical and Physical Sciences*, 310(1512), 347–363.

Cotton-Barratt, O. and Ord, T. (2015) "Existential risk and existential hope: Definitions." *Future of Humanity Institute: Technical Report 1.2015.*

Crisp, R. (2003) "Equality, priority, and compassion." *Ethics*, 113(4), 745-763.

Dyson, F. (1999) *The sun, the genome, and the internet*. Oxford: Oxford University Press.

Earman, J. (1992) Bayes or bust? A critical examination of Bayesian confirmation theory, Cambridge, MA: MIT Press.

Eder, A. (forthcoming). Evidential probabilities and credences. Forthcoming in *The British Journal for the Philosophy of Science*.

Elga, A. (2010) "Subjective probabilities should be sharp." *Philosophers' Imprint* 10.

Ellsberg, D. (2017) *The doomsday machine: Confessions of a nuclear war planner.* Bloomsbury Publishing USA.

Glymour, Clark N. (1980) "Why I Am Not a Bayesian", in his Theory and Evidence, Princeton, NJ: Princeton University Press.

Glynn, L. (2010) "Deterministic chance." *The British Journal for the Philosophy of Science* 61(1):51-80.

Greaves, H. (2017) "Population axiology." *Philosophy Compass* 12.11 (2017)

Greaves, H. and MacAskill, W. (2024) "The case for strong longtermism." Forthcoming in Greaves, H., Barrett, J. and Thorstad, D. (eds.), *Longtermism: Present action for the distant future*, Oxford University Press.

Gustafsson, J. (MS) "Against negative utilitarianism." Available online at https://johanegustafsson.net/papers/against-negative-utilitarianism.pdf. Accessed 11 September 2023.

Hanson, R. (1998) "The great filter - Are we almost past it?" Available at http://mason.gmu.edu/~rhanson/greatfilter.html (accessed 8 September 2023).

Häggström, O. (2016) *Here be dragons: Science, technology and the future of humanity*. Oxford University Press.

Kagan, S. (2019). *How to count animals, more or less*. Oxford, UK: Oxford University Press.

Kaku, M. (2018). *The Future of Humanity: Terraforming Mars, Interstellar Travel, Immortality, and Our Destiny Beyond Earth*. Doubleday.

Kolodny, N., and MacFarlane, J. "Ifs and oughts." *The Journal of Philosophy* 107(3):115-143.

Kopec, M. and Titelbaum, M. (2016). "The Uniqueness Thesis." *Philosophy Compass* 11 (4):189-200.

Lewis, D. (1979). "Counterfactual Dependence and Time's Arrow." *Noûs* 13(4):455–476.

Lewis, D. (1981) "Causal decision theory." *Australasian Journal of Philosophy* 59(1):5-30.

May, T. (2018). "Would Human Extinction Be a Tragedy?" *The New York Times*. December 17, 2018.

McMahan, J. (2002) *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford Academic.

Millett, P. and Snyder-Beattie, A. (2017) "Existential risk and cost-effective biosecurity." *Health Security* 15(4):373-83.

Mogensen, A. (forthcoming) "The weight of suffering". Forthcoming in *The Journal of Philosophy*.

Mokyr, J. (2017) *A Culture of Growth: The Origins of the Modern Economy*, Princeton: Princeton University Press.

Muehlhauser, L., and Bostrom, N. (2014). "Why We Need Friendly AI." *Think*, 13, pp. 41-47.

Nozick, R. (1969). "Newcomb's Problem and Two Principles of Choice." In: Rescher, N. (ed.) *Essays in Honor of Carl G. Hempel*. Synthese Library, vol 24. Springer, Dordrecht.

Ord, T. (2020) *The precipice: Existential risk and the future of humanity*. Hachette Books.

Peterson, M. (2017). *An Introduction to Decision Theory* (2nd ed.). Cambridge: Cambridge University Press.

Posner, R. (2004) *Catastrophe: Risk and response*. Oxford University Press.

Rees, M. (2003) *Our final century: Will the human race survive the twenty-first century?* Basic Books.

Riedener, S. (2022). "Human extinction from a Thomist perspective". In *Effective altruism and religion: Symmetries, tension, dialogue,* S. Riedener, D. Roser and M. Huppernbauer (eds.), Baden-Baden, Germany: pp. 187-210 (2022).

Robock, A., Oman, L. and Stenchikov, G. (2007) “Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences.” *Journal of Geophysical research: Atmospheres*, 112(D13107), doi:10.1029/2006JD008235.

Sagan, C. (1983) “Nuclear war and climate catastrophe: Some policy implications.” *Foreign Affairs* 62(2) (Winter 1983), pp.257-92.

Salow, B. (2018). “The Externalist’s Guide to Fishing for Compliments,” *Mind*, 127 (507), p. 691–728.

Schaffer, J. (2007) "Deterministic chance?" *The British Journal for the Philosophy of Science* 58 (2):113-140.

Scheffler, S. (2018) *Why worry about future generations?*. Oxford University Press.

Schoenfield, M. (2012). “Chilling out on epistemic rationality: A defense of imprecise credences (and other imprecise doxastic attitudes).” *Philosophical Studies* 158(2):197-219.

Schukraft, J. (2020) “Comparisons of capacity for welfare and moral status across species.” Available at https://rethinkpriorities.org/publications/comparisons-of-capacity-for-welfare-and-moral-status-across-species. Accessed 6 September 2023.

Sherwood, S. and Huber, M. (2010) “An adaptability limit to climate change due to heat stress”. *Proceedings of the National Academy of Sciences*, 107(21), 9552–5

Smith, C. and Davies, E. (2012) *Emigrating beyond earth: Human adaptation and space colonization*. Berlin: Springer-Verlag.

Snyder-Beattie, A. E., Sandberg, A., Drexler, K. E., & Bonsall, M. B. (2021). ”The timing of evolutionary transitions suggests intelligent life is rare.” *Astrobiology, 21*(3), 265-278. doi:10.1089/ast.2019.2149

Steele, K. (2024) “The minor role of totalism in the longtermists’ mathematics.” Forthcoming in Greaves, H., Barrett, J. and Thorstad, D., *Longtermism: Present action for the distant future*, Oxford University Press.

Thomas, T. (2022) "The asymmetry, uncertainty, and the long term." *Philosophy and Phenomenological Research*. Published online 14 November 2022. doi:10.1111/phpr.12927

Vallentyne, P. (2005). “Of mice and men: Equality and animals.” *The Journal of Ethics* 9(3-4):403-33.

White, R. (2010) "Evidential symmetry and mushy credence." *Oxford Studies in Epistemology* 3: 161-186.

Williamson, T. (2000) *Knowledge and its limits.* Oxford University Press.

Zubrin, R. (1999). *Entering Space: Creating a spacefaring civilization.* Tarcher/Putnam.