Existential Risk and Existential Hope: Definitions

Owen Cotton-Barratt* & Toby Ord†

We look at the strengths and weaknesses of two existing definitions of existential risk, and suggest a new definition based on expected value. This leads to a parallel concept: 'existential hope', the chance of something extremely good happening.

An existential risk is a chance of a terrible event occurring, such as an asteroid striking the earth and wiping out intelligent life – we could call such events *existential catastrophes*. In order to understand what should be thought of as an existential risk, it is necessary to understand what should be thought of as an existential catastrophe. This is harder than it first seems to pin down.

1. The simple definition

One fairly crisp approach is to draw the line at extinction:

> **Definition (i):** An existential catastrophe is an event which causes the end of existence of our descendants.

This has the virtue that it is a natural division and is easy to understand. And we certainly want to include all extinction events. But perhaps it doesn't cast a wide enough net.

> **Example A:** A totalitarian regime takes control of earth. It uses mass surveillance to prevent any rebellion, and there is no chance for escape. This regime persists for thousands of years, eventually collapsing when a supervolcano throws up enough ash that agriculture is prevented for decades, and no humans survive.

In Example A, clearly the eruption was bad, but the worst of the damage was done earlier. After the totalitarian regime was locked in, it was only a matter of time until something or other finished things off. We'd like to be able to talk about entering this regime as the existential catastrophe, rather than whatever event happens to end it. So we need another definition.

* Future of Humanity Institute, University of Oxford & Centre for Effective Altruism

† Future of Humanity Institute, University of Oxford

Although we'll now look at other definitions for existential catastrophes, we do like the simple definition. Luckily there's another term that's already understood: human extinction. Sometimes it's better to talk about extinction risks rather than existential risks, as 'existential risk' is a piece of jargon, whereas 'extinction risk' will be clear to everyone.

2. Bostrom's definition

Nick Bostrom introduced the concept of existential risks. He has defined them as follows:

> **Definition (ii)**: An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development.[1]

This definition deals well with Example A, placing the existential catastrophe at the point where the totalitarian regime arose, as this caused the *permanent and drastic destruction of [humanity's] potential for desirable future development*.

> **Example B**: A totalitarian regime takes control of the earth. There is only a slight chance that humanity will ever escape.

Is this an existential catastrophe? Bostrom's definition doesn't clearly specify whether it should be considered as one. Either answer leads to some strange conclusions. Saying it's not an existential catastrophe seems wrong as it's exactly the kind of thing that we should strive to avoid for the same reasons we wish to avoid existential catastrophes. Saying it is an existential catastrophe is very odd if humanity *does* escape and recover – then the loss of potential wasn't permanent after all.

The problem here is that potential isn't binary. Entering the regime certainly seems to *curtail* the potential, but not to eliminate it.

3. Definition via expectations

The idea that potential isn't binary motivates our suggested definition:

> **Definition (iii)**: An existential catastrophe is an event which causes the loss of a large fraction of expected value.

---

[1] Bostrom, Existential Risk Reduction as Global Priority, *Global Policy*, Vol 4, Issue 1 (2013), p15

This definition deals well with Example B. If we enter into the totalitarian regime and then at a later date the hope of escape is snuffed out, that represents *two* existential catastrophes under this definition. We lost most of the expected value when we entered the regime, and then lost most of the remaining expected value when the chance for escape disappeared.

A lot of the work of this definition is being done by the final couple of words. 'Value' refers simply to whatever it is we care about and want in the world, in the same way that 'desirable future development' worked in Bostrom's definition. And to talk about expectations we need to have some probabilities in mind. Here we are thinking of objective probabilities. Note that 'potential' in Bostrom's definition requires some similar work to making assumptions about probabilities.

4. Existential eucatastrophes and existential hope

If we enter the totalitarian regime and then manage to escape and recover, then we had an existential catastrophe which was balanced out by a subsequent gain in expected value. This kind of event gives us a concept parallel to that of an existential catastrophe:

> **Definition (iv)**: An *existential eucatastrophe*[2] is an event which causes there to be much more expected value after the event than before.

This concept is quite natural. We saw it in the context of escape from a regime which threatened the existence of a prosperous future. Our world has probably already seen at least one existential eucatastrophe: the origin of life. When life first arose, the expected value of the planet's future may have become much bigger. To the extent that they were not inevitable, the rise of multicellular life and intelligence may also have represented existential eucatastrophes.

In general successfully passing any 'great filter'[3] is an existential eucatastrophe, since beforehand the probability of passing it is small, so the expected value is much smaller than after the filter is dealt with.

---

[2] The word 'eucatastrophe' is made of the Greek root 'eu-' meaning 'good' and the word 'catastrophe' in its classical sense of a sudden turn. It was coined by Tolkien to refer to the sudden and unexpected turn for the better frequently found at the end of fairy tales. (Tolkien, John Ronald Reuel. *On fairy-stories*. Oxford University Press, 1947.)

[3] Hanson, Robin, *The Great Filter - Are We Almost Past It?*, 1998.

Armed with this concept, we can draw a new lesson. Just as we should strive to avoid existential catastrophes, we should also seek existential eucatastrophes.

In some ways, this isn't a new lesson at all. Under Bostrom's definition we are comparing ourselves to the most optimistic potential we could reach, so failing to achieve a eucatastrophe is itself a catastrophe. However we think more naturally in terms of events than non-events. If life fails to arise on a planet where it might have, it's much clearer to think of a failure to achieve a eucatastrophe than of an existential catastrophe stretching out over the billions of years in which life did not arise.

Just as we tend to talk about the existential risk rather than existential catastrophe, we want to be able to refer to the chance of an existential eucatastrophe; upside risk on a large scale. We could call such a chance an *existential hope*.

In fact, there are already people following both of the strategies this suggests. 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.

5. Conclusions

We were interested in pinning down what is meant by 'existential risk'. Much of the time, all of the definitions we've looked at will agree on whether something is an existential risk. Keeping it simple can be good, because it helps more people to understand. We therefore advocate talking about 'extinction risks' rather than 'existential risks' when the former term will work.

Nonetheless, we may sometimes have to consider more unusual scenarios. It's good to know how to make the definition work well there as it can help us to think about things more clearly. We think that the definition in terms of expectations does a better job of this than previous definitions.

In devising the notion of existential catastrophe (and hence existential risk) via expectations, we came across the dual concept we have called 'existential eucatastrophe' (and hence 'existential hope'). We think this captures a natural class of events, and what may be an important one. We hope that having a label for the concept may help others to make better judgements about what courses to pursue.