The following is a section from my MA Philosophy thesis *Uncertainty About the Expected Moral Value of the Long-Term Future: is Extinction Risk Reduction Valuable? ([link])* I'm sharing this section because I find it one of the most valuable and readable parts ot the thesis and is an easy excerpt. The arguments have been made before (Cf. Brauner & Grosse-Holz) and I learned them originally from people at the EA Foundation, but I hope a short and clear write-up is easier to reference.

I would not recommend readers to read the whole thesis, unless one plans to work on a (very) related topic. However, it may be interesting to skim through it.

Extinction risk reduction has little option value

Some people have suggested we should reduce existential risk for its *option value* (Bostrom, 2013; MacAskill, 2014). Bostrom (p. 24) writes:

> If we are indeed profoundly uncertain about our ultimate aims, then we should recognize that there is a great *option value* in preserving - and ideally improving - our ability to recognize value and to steer the future accordingly. Ensuring that there will be a future version of humanity with great powers and a propensity to use them wisely is plausibly the best way available to us to increase the probability that the future will contain a lot of value. To do this, we must prevent any existential catastrophe.

Remember that an existential catastrophe is "the extinction of Earth-originating intelligent life or the permanent and drastic failure of that life to realise its potential for desirable development" (Bostrom, 2013, p. 15). Since Bostrom includes all extinction events as existential catastrophes, I am focusing my criticism on the argument that reducing *extinction risk* has great option value. To criticize the argument, let me first deconstruct the argument into four premises and a conclusion.

Premise 1: We are profoundly uncertain about our ultimate aims.

Premise 2: If we are profoundly uncertain about our ultimate aims, then we should recognize that there is a great option value in preserving - and ideally

improving - our ability to recognize value and to steer the future accordingly.

From *Premise 1* and *2* follows:

Premise 3: There is great option value in preserving - and ideally improving - our ability to recognize value and to steer the future accordingly.

Premise 4: Preventing [extinction] preserves - and ideally improves - our ability to recognize value and to steer the future accordingly.

Conclusion: Preventing [extinction] has great option value.

Preventing extinction (and other existential catastrophes) probably ensures that there "will be a future version of humanity with great powers" (assuming technological development will continue). However, although preventing extinction is necessary to ensure that our descendants will have "a propensity to use [their great powers] wisely", it is not sufficient. We cannot assume that our descendants will necessarily be wise and altruistic without argument.[1] As a consequence, preventing extinction also leaves the option open that the future will contain a lot of negative value, because great power might be combined with a lack of wisdom or coordination. In what follows, I will criticize *Premise 2*: that when we are uncertain, there is great option value in preserving our ability to recognize value and steer the future accordingly.

How would 'preserving our ability to recognize and steer the future' yield option value? Normally, the option value of an asset is high when there is large uncertainty about the future need of the asset, and when losing the asset is irreversible (or comes with high costs). In this case, human civilization is the asset. Both conditions seem to be met; there is uncertainty about whether human civilization will be a positive influence on the value of the future, and extinction is mostly irreversible.[2] However, a third factor affecting option value

[1] In the above quote, Bostrom (2013) does not literally state that preventing existential catastrophe *ensures* that there will be a future version of humanity with great powers and a propensity to use them wisely, only that preventing existential catastrophe is *necessary*. However, he does not address the possibility of preventing existential catastrophe resulting resulting in an unwise future version of humanity anywhere in the paper.

[2] Given that Earth will remain hospitable to complex life for approximately a few hundred million to a billion years (O'Malley-James, J. T., Cockell, C. S., Greaves, J. S. and Raven, 2014), it is possible that another intelligent and complex civilization arises in that timespan. Thus, the capabilities lost by extinction of humanity are not irreversible for certain. On the other hand, extinction is not reversible in the sense that one can make a choice to reverse the situation based on new information.

is the extent to which one has the future ability to choose an option based on more information. This is where the argument is weakest.

Suppose we postpone extinction. Can future generations choose to change the course of the future if information is available that the expected value of the future is negative? Would humanity go as far as choosing extinction if the future looks bleak, as MacAskill (2014, p. 240) suggests humanity can?[3] Let's survey the possibilities. A future version of humanity is either capable or incapable to significantly change trajectory if it wants to[4], and either motivated or unmotivated to change trajectory of the expected moral value of the future looks negative. Below, in the left table, we see where option value resides: when humanity is motivated and able to change trajectory. In the right table we see where to expect the future to be negative.

Ability	**Motivation**	
	No	Yes
No		
Yes		*Option value*

Ability	**Motivation**	
	No	Yes
No	*Many possible negative futures*	*Many possible negative futures*
Yes	*Many possible negative futures*	*Few negative futures*

Figure 6a (left) and 6b (right). Possible combinations for a future version of humanity. 'Ability' stands for 'ability to significantly change the course of the future if they want to'. 'Motivation' stands for 'will want to significantly change the course of the future if it looks to have negative expected moral value'. Most of the option value resides in the scenarios in which the future looks very positive.

Only when humanity is both able and motivated to significantly change the course of the future do we have option value. However, suppose that our descendants both have the ability and the motivation to affect the future for the good of everyone, such that a future version of humanity is wise enough to recognize when the expected value of the future is negative and coordinated and powerful enough to go extinct or make other significant changes. As other authors have raised (Brauner & Grosse-Holz, 2018), given such a state of

[3] MacAskill (2014, p. 240) writes "If we continue to exist, then we always have the option of letting ourselves go extinct in the future (or, perhaps more realistically, of considerably reducing population size)."
[4] Ability is probably a combination of ability to alter the physical environment + ability to coordinate with other agents.

affairs it seems unlikely that the future would be bad! After all, humanity would be wise, powerful, and coordinated. Most of the bad futures we are worried about do not follow from such a version of humanity, but from a version that is powerful but unwise and/or uncoordinated.

To be clear, there would be a small amount of option value. There could be some fringe cases in which a wise and powerful future version of humanity would have good reason to expect the future to be better if they went extinct, and be able to do so. Or perhaps it would be possible for a small group of dedicated, altruistic agents to bring humanity to extinction, without risking even worse outcomes. At the same time they would need to be unable to improve humanity's trajectory significantly in any other way for extinction to be their highest priority. Furthermore, leaving open this option also works the other way around: a small group of ambitious individuals could make humanity go extinct if the future looks overwhelmingly positive.

In conclusion, deferring our choice to continue or not to our descendants yields little option value. In most of the scenarios in which they could decide to altruistically go extinct (or otherwise change the course of the future) it will not be needed, precisely because they would be altruistic and capable enough that the future would look bright and promising.