Allocating risk mitigation across time

Owen Cotton-Barratt*

This article is about priority-setting for work aiming to reduce existential risk. Its chief claim is that all else being equal we should prefer work earlier and prefer to work on risks that might come early. This is because we are uncertain about when we will have to face different risks, because we expect diminishing returns of extra work, and because we expect that more people will work on these risks in the future.

I explore this claim both qualitatively and with explicit models. I consider its implications for two questions: first, "When is it best to do different kinds of work?"; second, "Which risks should we focus on?".

As a major application, I look at the case of risk from artificial intelligence. The best strategies for reducing this risk depend on when the risk is coming. I argue that we may be underinvesting in scenarios where AI comes soon even though these scenarios are relatively unlikely, because we will not have time later to address them.

1. Overview

When we are very unsure about the difficulty of a problem, our subjective probability distribution tends to be distributed over several orders of magnitude, and we have diminishing marginal returns (in expectation) from extra work on the problem.

Suppose we are also unsure about when we may need the problem solved by. In scenarios where the solution is needed earlier, there is less time for us to collectively work on a solution, so there is less work on the problem than in scenarios where the solution is needed later. Given the diminishing returns on work, that means that a marginal unit of work has a bigger expected value in the case where the solution is needed earlier. This should update us towards working to address the early scenarios more than would be justified by looking purely at their impact and likelihood.

* Future of Humanity Institute, University of Oxford & Global Priorities Project

1.1. Timing of labour

In his article, The timing of labour aimed at existential risk, Ord lays out major considerations concerning what kind of work is best done when in reducing existential risk. These largely push towards meta-level work such as course-setting and growth in the short term, and object-level work later. However, uncertainty about when we need to be prepared to face risks adds some considerations in favour of doing more object-level work in the short term.

In the first place, there is simply a chance that we will not be around to face later risks, or that long-term preparations are ruined by unforeseen events. This corresponds to the catastrophe rate component of a discount rate, reducing the expected value of things whose payoff is further in the future.

Secondly, we today are in a privileged position with respect to scenarios where we must face risks early. Many people will be in a position to work on mitigating late risks, but only we are in a position to mitigate early risks. This means we could represent a much larger share of the total labour mitigating these risks than our share of the labour mitigating later risks, because there will be more time to work on later risks -- and perhaps more attention paid to them. If there are diminishing returns on extra labour on a risk, this increases the importance of object-level work soon. It is possible that this could overcome the extra leverage afforded by the meta-level approaches which have a longer payoff time.

1.2. How should we balance work on risks that will come at different times?

This consideration encourages us to do object-level work to reduce risks which can come at different times earlier than we would otherwise. But it also pushes us to prefer to work on risks which we might have to face early, over risks which we won't have to face until later.

In the remainder of this article I will focus on the specific question of how to compare work in AI safety which aims at scenarios where AI comes soon with scenarios where it comes later. This is both for the sake of concreteness and because it is an important question. Many of the considerations generalise to comparisons between work on mitigating other existential risks or groups of such risks.

2. Should AI safety work aim at scenarios where AI comes soon?

When it is developed, general-purpose artificial intelligence (AI) is likely to create big changes in the world. The possibility for this transition to go badly has led some to argue that AI safety is potentially a very valuable field. For this article I will assume that it is of high expected value, and consider the question: what is the right

distribution of resources within this field? In particular, how should we balance safety work which is aimed at scenarios where AI comes relatively soon with safety work which assumes that AI will not come for decades?

Of course the question of when we will have AI is something of a continuum, but to make the analysis simpler I will split it into scenarios where AI comes *soon*, meaning within the next two decades, and scenarios where AI comes *later*, meaning in the five or so decades after that. It is of course a distinct possibility that we will not get AI in this timeframe at all, but in that case current work is likely less important, so I'll omit it from the analysis.

The work you might do to improve safety conditional on AI coming soon may be quite different from work you'd do conditional on it coming later. Ord has argued that AI later favours meta-level work, which might include encouraging more people to consider careers in AI safety. But this could take decades to pay off. By contrast, AI soon favours object-level work such perhaps as looking for solutions to the value-loading problem. It may also push towards making assumptions about the nature of the problem or the nature of the AI that will arise, so that we have some cases solved (rather than a more comprehensive approach which might be preferable if we have more time).

In any case it needn't be that we should focus purely on the soon or purely on the later. But this may push us towards having a portfolio of separate efforts aimed at these different scenarios, rather than forgetting about one of them or looking for a single strategy which simultaneously handles them both well.

Note that the choice of cut-off between soon and later is somewhat arbitrary, based on an impression of where the strategies might naturally bifurcate. If you think a different cut-off is more appropriate you can change this without affecting the structure of the analysis.

2.1. Major Considerations

Now, almost everyone agrees that it is much more likely that AI will come later than come soon. Isn't this a good reason to focus on the work that helps if AI comes later? Not necessarily. This does provide a significant factor in favour of such work, but it's possible that that this could be outweighed by other factors.

There are two major factors which seem to push towards preferring more work which focuses on scenarios where AI comes soon. The first is nearsightedness: we simply have a better idea of what will be useful in these scenarios. The second is diminishing marginal returns: the expected effect of an extra year of work on a problem tends to decline when it is being added to a larger total. And because there is a much larger time horizon in which to solve it (and in a wealthier world), the

problem of AI safety when AI comes later may receive many times as much work as the problem of AI safety for AI that comes soon. On the other hand one more factor preferring work on scenarios where AI comes later is the ability to pursue more leveraged strategies which eschew object-level work today in favour of generating (hopefully) more object-level work later.

In order to compare the size of these effects with the higher likelihood of getting AI later, I introduce explicit models in Section 3. My main contribution is the functional form of these models, but to see roughly which direction they push in I have gathered some estimates of the model parameters. Overall these models push me towards thinking that we should take seriously the idea that we may be relatively under-investing in scenarios where AI comes soon.

2.2. Alternative perspectives

Because explicit models can be brittle, it is helpful to try other routes to answer our main questions. Here another tack is to ask: without explicit correction, should we expect to under- or over-invest in scenarios where AI comes soon? I will give a brief analysis, and can see a case in either direction.

On the one hand, we often like things to be close and concrete. Perhaps this means that a concern for risks would translate primarily to work on safety for AI-soon scenarios.

On the other hand, people hate being wrong and looking silly. Although it may be appropriate to have a subjective probability today of say 1% that AI comes soon, in most of the 99% of cases where it doesn't come soon, it will seem in retrospect with better understanding that the probability was very much lower than 1%, and perhaps effectively zero. It is hard to ask someone to spend twenty years of their life working on something that will very likely look like it was a waste of time afterwards!

On balance I think this effect may be stronger in pushing people away from doing work focusing on scenarios where it come soon. This agrees with my tentative conclusions from the models, and makes me think we should seek more work which assumes AI comes soon.

Is there a danger to such work? Perhaps: if it becomes too large and communicates a confidence that AI is coming soon, then it may look silly when the threat doesn't materialise on schedule. This could in turn make it harder to get attention for the more likely scenario of AI later. However while I do think it's appropriate to worry about this, it should be possible to guard against it.

3. Models

Modelling the value of any work to reduce existential risk is hard, because we don't have a good sense of how existential risk reduction should be traded off against other more normal-looking goods (which may indirectly impact existential risk). To sidestep this problem, the models I will consider just make comparisons between different kinds of existential risk reduction work. This means that the benefits in our benefit:cost ratios will share units, and admit easier comparisons. In the appendix I'll present the full models I've produced so far. However they include several variables which likely don't change the answer much, so for usability I give simplified versions of the models in this section.

All of the models here just produce estimates of the broad value of working on different areas. In order to make comparisons between opportunities in different areas, one should also estimate the leverage ratios for those opportunities. And of course these models are still quite crude, and I'd be hesitant to take their output at face value.

3.1. First model – direct value of work

In our first model we produce a framework for comparing the direct value of a little extra work on the AI-soon safety problem with a little extra work on the AI-later safety problem. We make the assumption that in either case work we do has no effect on the total (relevance-adjusted) amount of other work done on the problems. I don't think this assumption is accurate, and I look at relaxing it in the next model. I do think it's informative to consider the model with this assumption.

First some notation for the model:

- S denotes the AI-soon safety problem; that is that we first have to deal with roughly human-level artificial general intelligence in the next 20 years.
- L denotes the AI-later safety problem; that is that we first have to deal with roughly human-level artificial general intelligence 20-70 years from now.
- $p(\mathrm{X})$ denotes the probability that we will face problem X.
- $m(\mathrm{X})$ denotes the probability that a marginal unit of work on X will solve X.
- $v(\mathrm{X})$ denotes the expected value of a marginal unit of work on X, arising from its chance of averting an existential catastrophe.

Now we can factor the value of extra work on these problems:

$$v(S) = p(S) \times m(S)$$
$$v(L) = p(L) \times m(L)$$

To compare them we can consider the ratio:

$$\frac{v(S)}{v(L)} = \frac{p(S)}{p(L)} \times \frac{m(S)}{m(L)}$$

Even though S and L are not crisply defined, this is getting towards something we can provide crude estimates for. My approach was to ask directly for estimates of $p(S)$ and $p(L)$, and to apply a further model to estimate $m(S)$ and $m(L)$.

Both S and L seem to be problems where we have very little idea how difficult they are. In this context, I think we should expect our chances of success to be very approximately linear with the logarithm of the resources devoted to them. This means that the marginal chance of success is proportional to 1/x, where x is the total amount of resources that will be devoted to the problem before the point where we need a solution.

Let $r(S)$ and $r(L)$ denote the total amount of relevance-weighted resources that will be devoted towards S and L, in the worlds where AI comes soon or later respectively. With some minor simplification from the model presented in the appendix, we can use this to model:

- $m(S) = \frac{n(S)}{r(S)}$
- $m(L) = \frac{n(L)}{r(L)}$

where $n(S)$ and $n(L)$ measure the relevance of an extra unit of work on the problem now (accounting for nearsightedness).

That gives the ratio of value of soon work to later work as:

$$\frac{v(S)}{v(L)} = \frac{p(S)}{p(L)} \times \frac{n(S)}{n(L)} \times \frac{r(L)}{r(S)}$$

To recap, we have three comparative terms:

1. $\frac{p(S)}{p(L)}$ expresses the ratio between the likelihood of getting AI in the soon period compared to the later period.

2. $\frac{n(S)}{n(L)}$ is the ratios of the nearsightedness factors for the work: how suboptimal is our work on S (compared to when we have nearly reached AI, if it comes soon), compared to how suboptimal is our work on L (with the same comparison).

3. $\frac{r(L)}{r(S)}$ expresses the ratio between the total resources that each problem will receive in total, conditional on it being the relevant one. We might denominate these resources in (relevance adjusted) dollars or researcher-years. It is perhaps better thought of as $\frac{1}{r(S)}/\frac{1}{r(L)}$, because we really want to use the ratio of expectations of $\frac{1}{r(S)}$ and $\frac{1}{r(L)}$.

You might like to take a few moments to think about your personal estimates for each of these components. I have collected a few estimates which I present in Section 3.3.

3.2. Second model – promoting area growth

The particularly questionable assumption in the first model was that work today has no influence on the quantity of future work. For the second model, we consider a different extreme where there is a strong feedback effect where extra work today produces extra future work.

In particular we will assume that the total work done on each of S and L grows exponentially. So if we intervene so as to increase the total work that has ever been done on S by 1%, the total amount of work that has been done on S will continue to be 1% higher than it otherwise would have been. This model assumes that growth of attention to the problems is largely endogenous, driven by attention and work they have already received. It also assumes that work on each of them will continue to grow exponentially until the time when it might be needed, rather than levelling off after reaching some maximum. These assumptions are probably unrealistic, and they are pushing in favour of extra work on L compared to S. We also assume that work on S and L are independent and will not drive each other; this is also unrealistic but I am not sure which direction it pushes the answer in.

We will continue to use the same model of value of extra work on the problems. Since this says it is logarithmic with the work done, and the work done increases exponentially, this means that the expected benefit of a boost to one problem doesn't depend on the date at which we need a solution.

Then with this model the value of soon work to later work is:

$$\frac{v(S)}{v(L)} = \frac{p(S)}{p(L)} \times \frac{h(L)}{h(S)}$$

Here the first term is the ratio of probabilities of facing the problems, as in the previous model. The second term is:

4. $\frac{h(L)}{h(S)} = \frac{1}{h(S)} / \frac{1}{h(L)}$ is the ratio of the total historical work on the two problems. Note that *h(X)* measures the work on *X* done up till now, while *r(X)* includes both past and future work. Again we can denominate this in any appropriate unit.

3.3. Estimates of model parameters

In order to apply these simple models, we need estimates for the parameters. I have a small collection of estimates of $\frac{p(S)}{p(L)}$, $\frac{n(S)}{n(L)}$, $\frac{r(L)}{r(S)}$, and $\frac{h(L)}{h(S)}$ from researchers at the Future of Humanity Institute:

- $\frac{p(S)}{p(L)}$ has eight estimates following a discussion. The median of these is **0.22**, and the range is 0.17 – 1.7.
- $\frac{n(S)}{n(L)}$ has five estimates without discussion. They are 0.5, 1, **1.5**, 2, 3.
- $\frac{r(L)}{r(S)}$ has five estimates without discussion. They are 1.4, 5, **10**, 20, 30.
- $\frac{h(L)}{h(S)}$ has five estimates without discussion. They are 1, 1.5, **3**, 6.5, 10.

Using the bolded median estimates, the first model gives $\frac{v(S)}{v(L)} = 0.22 \times 1.5 \times 10 = 3.3$; *i.e.* extra resources on the soon problem are around three times more valuable (for their direct effect) as ones on the later problem. The second model gives for these estimates $\frac{v(S)}{v(L)} = 0.22 \times 3 = 0.66$; *i.e.* that counting just growth effects, work on the later problem is perhaps half again as valuable as work on the soon problem.

3.4. Conclusions from models

The two models presented here deliberately try to err in different directions. I think the true answer is likely to involve components of each and lie somewhere in-between. As well as uncertainty about how appropriate the models are, there is substantial uncertainty about model parameters.

In spite of this uncertainty, I think that we can draw some useful conclusions from the models if we're willing to make estimates of the parameters..

First, they provide solid reason to think that working on safety in the AI-soon and AI-later scenarios are of similar value. That is, neither beats the other by several orders of magnitude. This means that both are likely worth bearing in mind: a high-leverage opportunity in one is often better in expectation than a typical opportunity in the other.

Second, they support the idea that work on the AI-later problem is primarily useful by attracting attention and more work, and should be significantly optimised for this. On the other hand it is less clear how this compares to the direct value of the work for AI-soon scenarios.

Third, they weakly suggest that we should increase our focus on AI-soon scenarios. While I do not take the model outputs as conclusive evidence that it is better to target these scenarios, I do think that they demonstrate that it is a serious possibility, and that the lack of attention they have received and shorter time-horizon for dealing with them in may overcome the fact that they are less likely.

Fourth, they suggest there could be strategic value in putting more work into obtaining estimates of the model parameters. There was a lot of variance in individual estimates; enough that changing from one person's estimates to another might tip the conclusion from definitely preferring soon-focused work to definitely preferring later-focused work. And these parameters have mostly not received much previous attention, so it may be easy to tighten the estimates by sharing governing considerations. In a similar vein it could be valuable to explore the conclusions of variations on these models, and more generally they may be worth applying to other existential risks.

Acknowledgements: I am particularly grateful to Daniel Dewey and Toby Ord for conversations on this topic and comments on earlier drafts. Thanks also to Stuart Armstrong, Nick Bostrom, Eric Drexler, Seb Farquhar, Anders Sandberg, and Carl Shulman for, variously: comments, conversation, and parameter estimates.

Appendix – full models

When presenting the two models above, in the interest of simplicity I omitted some variables, where it seemed like the ratio between them would be close to 1. In these appendices, I give expanded forms which more comprehensively cover the factors which may differ between work on different risks.

A. First model – direct value of work

In our first model we produce a framework for comparing the direct value of a little extra work on mitigating two different risks. We make the assumption that in either case work we do has no effect on the total (relevance-adjusted) amount of other work done on the problems. I don't think this assumption is accurate, and I look at relaxing it in the next model. I do think it's informative to consider the model with this assumption.

First some notation for the model:

- A and B denote two existential risks where we are interested in comparing between the expected value of extra work on the two.
- $p(X)$ denotes the probability that we will face problem X. Note that this is meant to be an absolute probability, not conditional on getting to the the point where we might face X.
- $m(X)$ denotes the probability that a marginal unit of work on X will solve X.
- $v(X)$ denotes the expected value of a marginal unit of work on X, arising from its chance of averting an existential catastrophe.

Now we can factor the value of extra work on these problems:

$$v(A) = p(A) \times m(A)$$
$$v(B) = p(B) \times m(B)$$

To compare them we can consider the ratio[1]:

$$\frac{v(A)}{v(B)} = \frac{p(A)}{p(B)} \times \frac{m(A)}{m(B)}$$

We will then apply a further model to estimate $m(A)$ and $m(B)$.

Many possible A and B seem to be problems where we have very little idea how difficult they are. This model will be appropriate in that case. If we do have a better idea of the difficulty, we should model that directly. If we *are* very uncertain, I think

[1] Note that this ratio isn't something that we should consider expectations of. We care about the ratio of expectations, not the expectation of the ratio.

we should expect our chances of success to be very approximately linear with the logarithm of the resources devoted to them. This means that the marginal chance of success is proportional to 1/x, where x is the total amount of resources that will be devoted to the problem before the point where we need a solution.

Let $r(A)$ and $r(B)$ denote the total amount of relevance-weighted resources that will be devoted towards A and B, in the worlds where we ultimately have to face those risks. Then we can use this to model:

- $m(A) = \frac{c(A) \times n(A)}{r(A)}$
- $m(B) = \frac{c(B) \times n(B)}{r(B)}$

where $c(A)$ and $c(B)$ are the constants accounting for the fact that we may have different ideas about how hard A and B are, and $n(A)$ and $n(B)$ measure the relevance of an extra unit of work on the problem now (accounting for nearsightedness).

That gives the ratio of value of soon work to later work as:

$$\frac{v(A)}{v(B)} = \frac{p(A)}{p(B)} \times \frac{c(A)}{c(B)} \times \frac{n(A)}{n(B)} \times \frac{r(B)}{r(A)}$$

To recap, we have four comparative terms:

1. $\frac{p(A)}{p(B)}$ expresses the ratio between the likelihood of having to face the two risks.

2. $\frac{c(A)}{c(B)}$ is the hardest term to unpack; it comes from the model of problem difficulty and is a measure of how difficult we think the problems may be. Luckily, it isn't typically that large: it measures the number of orders of magnitude we think the difficulty may be spread over, so even if we thought it spread over 6 orders of magnitude for one problem and only 3 for the other, the ratio could only reach 2. For these reasons I omitted it from the simple model.

3. $\frac{n(A)}{n(B)}$ is the ratios of the nearsightedness factors for the work: how suboptimal is our work on A for nearsighteness, compared to how suboptimal is our work on B.

4. $\frac{r(B)}{r(A)}$ expresses the ratio between the total resources that each problem will receive in total, conditional on it being relevant. We might denominate these resources in (relevance adjusted) dollars or researcher-years. It is perhaps

better thought of as $\frac{1}{r(A)}/\frac{1}{r(B)}$, because we really want to use the ratio of expectations of $\frac{1}{r(A)}$ and $\frac{1}{r(B)}$.

B. Second model – promoting area growth

The particularly questionable assumption in the first model was that work today has no influence on the quantity of future work. For the second model, we consider a different extreme where there is a strong feedback effect where extra work today produces extra future work.

In particular we will assume that the total work done on each of A and B grows exponentially. So if we intervene so as to increase the total work that has ever been done on A by 1%, the total amount of work that has been done on A will continue to be 1% higher than it otherwise would have been. This model assumes that growth of attention to the problems is largely endogenous, driven by attention and work they have already received. It also assumes that work on each of them will continue to grow exponentially until the time when it might be needed, rather than levelling off after reaching some maximum.

We will continue to use the same model of value of extra work on the problems. Since this says it is logarithmic with the work done, and the work done increases exponentially, this means that the expected benefit of a boost to one problem doesn't depend on the date at which we need a solution.

Then with this model the value of soon work to later work is:

$$\frac{v(A)}{v(B)} = \frac{p(A)}{p(B)} \times \frac{c(A)}{c(B)} \times \frac{g(A)}{g(B)} \times \frac{h(B)}{h(A)}$$

Here the first two terms are the same as in the previous model. The final two terms are:

1. $\frac{g(A)}{g(B)}$ is the ratio of the exponential growth rates the two problems will receive. If we thought one was easier to build support for we might think the rate of endogenous growth was higher. It is unclear which is higher and the ratio is unlikely to be large, so I omitted this from the simple model.

2. $\frac{h(B)}{h(A)} = \frac{1}{h(A)}/\frac{1}{h(B)}$ is the ratio of the total historical work on the two problems. Again we can denominate this in any appropriate unit.

Note that the assumption about logarithmic returns was doing quite a bit of work in these models. If you think the returns diminish at a different rate, that could substantially affect the model conclusions.