# Crucial Considerations and Wise Philanthropy

Nick Bostrom
Oxford University
www.nickbostrom.com

[Originally transcribed and published by Pablo Stafforini. On July 9th, 2014, Nick Bostrom gave a talk on "Crucial Considerations and Wise Philanthropy" (audio|slides) at Good Done Right, a conference on effective altruism held at All Souls College, Oxford. I found the talk so valuable that I decided to transcribe it.]

Formats: HTML, PDF

---

This talk will build on some of the ideas that Nick Beckstead was talking about before lunch. By contrast with his presentation, though, this will not be a well-presented presentation.

[laughter]

This is very much a work in progress, and so there's going to be some jump cuts, and some of the bits will be muddled, etc. But I'll look forward to the discussion part of this.

## What is a crucial consideration?

So I want to talk about this concept of a *crucial consideration*, which comes up in the work that we're doing a lot. Suppose you're out in the forest and you have a map and a compass, and you're trying to find some destination. You're carrying some weight, maybe you have a lot of water because you need to hydrate yourself to reach your goal and carry weight, and trying to fine-tune the exact direction you're going. You're trying to figure out how much water you can pour out, to lighten your load without having too little to reach your destination.

All of these are normal considerations: you're fine-tuning the way you're going to make more rapid progress towards your goal. But then you look more closely at this compass that you have been using, and you realize that the magnet part has actually come loose. This means that the needle might now be pointing in

a completely different direction that bears no relation to North: it might have rotated some unknown number of laps or parts of a lap.

With this discovery, you now completely lose confidence in all the earlier reasoning that was based on trying to get the more accurate reading of where the needle was pointing. This would be an example of a crucial consideration in the context of orienteering. The idea is that there could be similar types of consideration in more important contexts, that throw us off completely. So a crucial consideration is a consideration such that if it were taken into account it would overturn the conclusions we would otherwise reach about how we should direct our efforts, or an idea or argument that might possibly reveal the need not just for some minor course adjustment in our practical endeavors but a major change of direction or priority.

Within a utilitarian context, one can perhaps try to explicate it as follows: a crucial consideration is a consideration that radically changes the expected value of pursuing some high-level subgoal. The idea here is that you have some evaluation standard that is fixed, and you form some overall plan to achieve some high-level subgoal. This is your idea of how to maximize this evaluation standard. A crucial consideration, then, would be a consideration that radically changes the expected value of achieving this subgoal, and we will see some examples of this. Now if you widen the context not limited to some utilitarian context, then you might want to retreat to these earlier more informal formulations, because one of the things that could be questioned is utilitarianism itself. But for most of this talk we will be kind of thinking about that component.

There are some related concepts that are useful to have. So a *crucial consideration component* will be an argument, idea or datum which, while not on its own amounting to a crucial consideration, seems to have a substantial probability of maybe being able to serve a central role within a crucial consideration. It's the kind of thing [of which we would say:] "This looks really intriguing, this could be important; I'm not really sure what to make of it at the moment." On its own maybe it doesn't tell us anything, but maybe there's another piece that, when combined, will somehow yield an important result. So those kinds of crucial consideration components could be useful to discover.

Then there's the concept of a *deliberation ladder*, which would be a sequence of crucial considerations, regarding the same high-level subgoal, where the considerations hold in opposing directions. Let's look at some examples of these kinds of crucial consideration ladders that help to illustrate the general predicament.

## Should I vote in the national election?

Let's take this question: "Should I vote in the national election?" At the sort of "level one" of reasoning, you think, "Yes, I should vote to put a better candidate in office." That clearly makes sense.

Then you reflect some more: "But, my vote is extremely unlikely to make a

difference. I should not vote, but put my time to better use."

(These examples are meant to illustrate the general idea; it's not so much I want a big discussion as to these particular examples, they're kind of complicated. But I think they will serve to illustrate the general phenomenon.)

So here we have gone from "Yes, we should vote," making a plan to get to the polling booth, etc. And then, with the consideration number two, "No, I should not vote. I should do something completely different."

Then you think, "Well, although it's unlikely that my vote will make a difference, the stakes are very high: millions of lives are affected by the president. So even if the chance that my vote will be decisive is one in several million, the expected benefit is still large enough to be worth a trip to the polling station." So I just went back to the television, turned on the football game, and now it turns out I should vote, so we have a reversed direction.

Then you continue to think, "Well, if the election is not close, then my vote will make no difference. If the election *is* close, then approximately half of the votes will be for the wrong candidate, implying either that the candidates are exactly or almost exactly of the same merit, so it doesn't really matter who wins, or typical voters' judgment of the candidates' merits is extremely unreliable, and carries almost no signal, so I should not bother to vote."

Now you sink back into the comfy sofa and bring out the popcorn or whatever, and then you think, "Oh, well, of course I'm a much better judge of the candidates' merits than the typical voter, so I should vote."

Then you think, "Well, but psychological studies show that people who tend to be overconfident almost everybody believes themselves to be above average, but they are as likely to be wrong as right about that. If I am as likely to vote for the wrong candidate as is the typical voter, then my vote would have negligible information to the selection process, and I should not vote."

Then we go on…

[laughter]

"Okay, I've gone through all of this reasoning that really means that I'm special, so I should vote."

But then, "Well, if I'm so special, then the opportunity cost…"

[laughter]

(This is why I warned you all against becoming philosophers.)

[laughter]

So, I should do something more important. But if I don't vote my acquaintances will see that I have failed to support the candidates that we all think are best, they would think me weird and strange, and disloyal. Then that would maybe

diminish my influence, which I could otherwise have used for good ends, so I should vote after all.

But it's important to stand up for one's convictions, to stimulate fruitful discussion. They might think me like really sophisticated if I explained all this complicated reasoning for voting, and that might increase my influence, which I can then invest in some good cause. Etc, etc, etc.

There is no reason to think that the ladder would stop there; it's just that we ran out of steam at this point. If you end at some point, you might then wonder, maybe there are further steps on the ladder, and how much reason do you really think you have for the conclusion you're temporarily at, at that stage?

## Should we favor more funding for x-risk tech research?

I want to look at one other example of a deliberation ladder more in the context of technology policy and and x-risk. This is a kind of argument that can be run with regard to certain types of technologies, whether we should try to promote them or get more funding from.

The technology here is nanotechnology—this is in fact the example where this line of reasoning originally came up. Some parts of this hark back to Eric Drexler's book *Engines of Creation*, where he actually advocated this line of thinking [ch. 12]. So we should fund nanotechnology—this is the "level one" reasoning— because there are many potential future applications: medicine, manufacturing, clean energy, etc. It would be really great if we had all those benefits.

But it also looks like nanotechnology could have important military applications, and it could be used by terrorists etc., to create new weapons of mass destruction that could pose a major existential threat. If it's so dangerous, no, maybe we shouldn't really fund it.

But if this kind of technology is possible, it will almost certainly be developed sooner or later, whether or not we decide to pursue it. ('We' being maybe the people in this room or the people in Britain or Western democracies.) If responsible people refrain from developing it, then it will be developed by irresponsible people, which would make the risks even greater, so we should fund it.

(You can see that the same template could be relevant for evaluating other technologies with upsides and downsides, besides nanotechnology.)

But we are already ahead in its development, so extra funding would only get us there sooner, leaving us less time to prepare for the dangers. So we should not add funding: the responsible people can get there first even without adding funding to this endeavor.

But then you look around and see virtually no serious effort to prepare for the dangers of nanotechnology—and this is basically Drexler's point back in *Engines*—, because serious preparation will begin only after a massive project

is already underway to develop nanotechnology. Only then will people take the prospect seriously.

The earlier a serious Manhattan-like project to develop nanotechnology is initiated, the longer it will take to complete, because the earlier you start, the lower the foundation from which you begin. The actual project will then run for longer, and that will then mean more time for preparation: serious preparation only starts when the project starts, and the sooner the project starts, the longer it will take, so the longer the preparation time will be. And that suggests that we should push as hard as we can to get this product launched immediately, to maximize time for preparation.

But there are more considerations that should be taken into account. The level of risk will be affected by factors other than the amount of serious preparation that has been made, specifically to counter the threat from nanotechnology. For instance, machine intelligence or ubiquitous surveillance might be developed before nanotechnology, eliminating or mitigating the risks of the latter. Although these other technologies may pose great risks of their own, those risks would have to be faced anyway. And there's a lot more that can be said.

Nanotechnology would not really reduce these other risks, like the risks from AI, for example. The preferred sequence is that we get superintelligence or ubiquitous surveillance before nanotechnology, and so we should oppose extra funding for nanotechnology even though superintelligence and ubiquitous surveillance might be very dangerous on their own, including posing existential risk, given certain background assumptions about the technological completion conjecture—that in the fullness of time, unless civilization collapses, all possible general useful technologies will be developed—, these dangers will have to be confronted, and all our choice really concerns is the sequence in which we confront them. And it's better to confront superintelligence before nanotechnology because superintelligence can obviate the nanotechnology risk, but not *vice versa*.

However, if people oppose extra funding for nanotechnology, then people working in nanotechnology will dislike those people who are opposing it. (This is also point from Drexler's book.) But other scientists might regard these people who oppose funding for nanotechnology as being anti-science and this will reduce our ability to work with these scientists, hampering our efforts on more specific issues—efforts that stand a better chance of making a material difference to any attempt on our part to influence the level of national funding for nanotechnology. So we should not oppose nanotechnology. That is, rather than opposing nanotechnology—we may try to slow it down a little bit, but we are a small group and we can't make a big difference—, we should work with the nanotechnology scientists, be their friend, and then maybe try to influence on the margin, so that they develop nanotechnology in a slightly different way or add some safeguards, and stuff like that.

Again, there is no clear reason to think that we have reached the limit of the

level of deliberation that we could apply to this. It's disconcerting because it looks like the practical upshot keeps switching back and forth as we look more deeply into the search tree, and we might wonder why this is so. I think that these deliberation ladders are particularly likely to turn up when one is trying to be a thoroughgoing utilitarian and one really takes the big-picture question seriously.

## Crucial considerations and utilitarianism

[Let's consider] some possible reasons for why that might be. If we compare, for example, the domain of application of utilitarianism to another domain of application, say if you have an ordinary human preference function—you want a flourishing life, like a healthy family, a successful career and some relaxation, like a typical human values—if you're trying to satisfy those, it looks less likely that you will encounter a large number of these crucial considerations. Why might that be?

One possible explanation is that we have more knowledge and experience of human life at the personal level. Billions of people have tried to maximize an ordinary human utility function and have received a lot of feedback and a lot of things have been tried out. So we already know some of the basics like, if you want to go on for decades, it's a good idea to eat, things like that.

[laughter]

They're not something we need to discover. And maybe our preferences in the first place have been shaped to more or less fit the kind of opportunities we can cognitively exploit in the environment by evolution. So we might not have some weird preference that there was no way that we could systematically satisfy. Whereas with utilitarianism, the utilitarian preference extends far and wide beyond our familiar environment, including into the cosmic commons and billions of years into the future and super advanced civilizations: what they do matters from the utilitarian perspective, and matters a lot. Most of what the utilitarian preference cares about is stuff that we have no familiarity with.

Another possible source of crucial considerations with regard to utilitarianism is difficulties in understanding the goal itself. For example, if one tries to think about how to apply utilitarianism to a world that has a finite probability of being infinite, [one will] run into difficulties in terms of how to measure different infinite magnitudes and still seeing how we could possibly make any difference to it. I have a big paper about that and we don't need to go into that. There are some other issues that consist in actually trying to articulate utilitarianism to deal with all these possible cases.

The third possible reason here is that one might think that we are kind of close, not super close, but close to some pivot point of history. That means that we might have special opportunities to influence the long-term future now. And we're still far enough away from this: it's not obvious what we should do to have

the maximally beneficial impact on the future. But still close enough that we can maybe begin to perceive some contours of the apparatus that will shape the future. For example, you may think that superintelligence might be this pivot point, or one of them (there may be x-risk pivot points as well), that we will confront in this century, then it might just be that we are barely just beginning to get the ability to think about those things, which introduces a whole set of new considerations that might be very important.

This could affect the personal domain as well. It's just like with an ordinary person's typical utility function: they probably don't place a million times more value on living for a billion years than living for a hundred years, or a thousand times more value on raising a thousand children than on raising one child. So even though the future still exists, it just doesn't weigh as heavily in a normal human utility function as it does for utilitarians.

[Fourthly,] one might also argue that we have recently discovered some key exploration tools that enable us to make these very important discoveries about how to be a good utilitarian. And we haven't yet run the course with these tools, so we keep turning up like fundamental new important discoveries using these exploration tools. That's why there seem to be so many crucial considerations being discovered. We might talk a little bit about some of those later in the presentation.

## Evaluation functions

Now let me come at this from a slightly different angle. In chess, the way you would ideally play is you would start by thinking the possible moves that you could make, then the possible responses that your component could make, and your responses to those responses. Ideally, you would think that through all the way to the end state, and then just try to select a first move that would be best from the point of view of winning when you could calculate through the entire game tree. But that's computational infeasible because the tree branch is too much: you have an exponential number of moves to consider.

So what you instead have to do is to calculate explicitly some number of plies ahead. Maybe a dozen plies ahead or something like that. At that point, your analysis has to stop, and what you do is to have some evaluation function which is relatively simple to compute, which tries to look at the board state that could result from this sequence of six moves and countermoves, and in some rough and ready way try to estimate how good that state is. A typical chess evaluation function might look something like this.

## Evaluation function

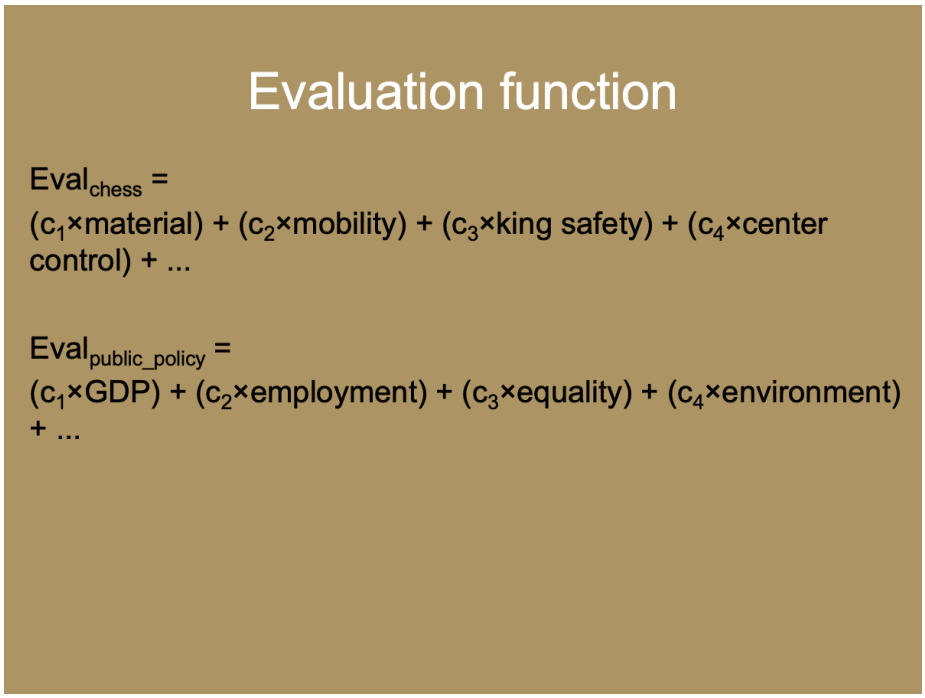
$\text{Eval}_{\text{chess}}$ =
$(c_1 \times \text{material}) + (c_2 \times \text{mobility}) + (c_3 \times \text{king safety}) + (c_4 \times \text{center control}) + \ldots$

You have some term that evaluates how much material we have, like having your queen and a lot of pieces is beneficial. The opponent having few of those is also beneficial. We have some metric like a pawn is worth one and queen is worth, I don't know, 11 or something like that.

So you weigh that up—that's one component in the evaluation function. Then maybe consider how mobile your pieces are. If they're all crammed in the corner, that's usually an unpromising situation, so you have some term for that. King safety… Center control adds a bit of value: if you control the middle of the board, we know from experience that tends to a good position.

So what you do is calculate explicitly some number of steps ahead and then you have this relatively unchanging evaluation function that is used to figure out which of these initial games that you could play would be resulting in the most beneficial situation for you. These evaluation functions are mainly derived from some human chess masters who have a lot of experience playing with the game. The parameters, like the weight you assign to these different features, might also be learned by machine intelligence.

We do something analogous to that in other domains. Like a typical traditional public policy, social welfare economists might think that you need to maximize some social welfare function which might takes a form like this.




GDP? Yes, we want more GDP, but we also have to take into account the amount of unemployment, maybe the amount of equality or inequality, some factor for the health of the environment. It might not be that whatever we write there s exactly the thing that is equivalent to moral goodness fundamentally considered. But we know that these things tend to be good, or we think so.

This is a useful approximation of true value that might be more tractable in a practical decision-making context. One thing I can ask, then, is if there is something similar to that for moral goodness. You want to do the morally best thing you can do, but to calculate all of these out from scratch just looks difficult or impossible to do in any one situation. You need more stable principles that you can use to evaluate different things you could do. Here we might look at the more restricted version of utilitarianism. We can wonder what we might put in there.

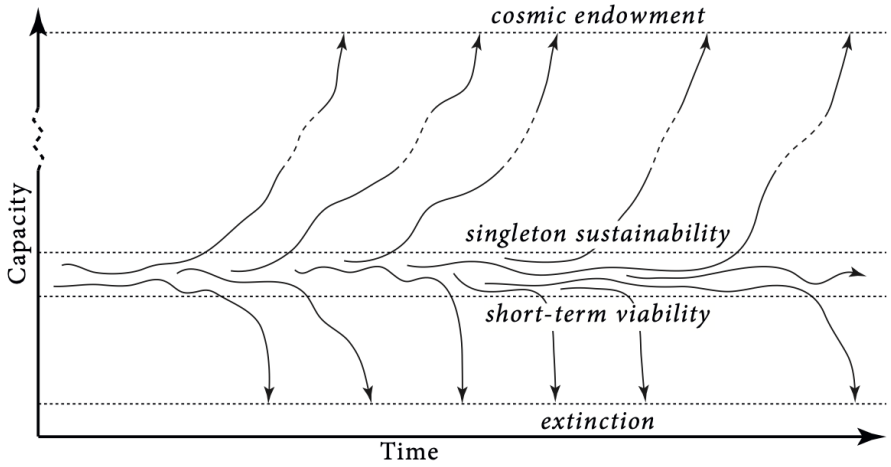
## Evaluation function

$\text{Eval}_{\text{chess}}$ =
$(c_1 \times \text{material}) + (c_2 \times \text{mobility}) + (c_3 \times \text{king safety}) + (c_4 \times \text{center control}) + \ldots$

$\text{Eval}_{\text{public\_policy}}$ =
$(c_1 \times \text{GDP}) + (c_2 \times \text{employment}) + (c_3 \times \text{equality}) + (c_4 \times \text{environment}) + \ldots$

$\text{Eval}_{\text{utilitarian}} = ?$

Here we can hark back to some of the things Beckstead talked about. If we plot capacity, which could be level of economic development and technological sophistication, stuff like that, on one axis and time on the other, my view is that the human condition is a kind of metastable region on this capability axis.




You might fluctuate inside for a while, but the longer the time scale you're considering, the greater the chance that you will exit that region in either the downwards direction and go extinct—if you have too few resources below the minimum viable population size, you go extinct (that's one attractor state: once you're extinct, you tend to stay extinct)—or in the upwards direction: we get through to technological maturity, start colonization process and the future of earth-originating intelligent life might just then be this bubble that expands at some significant fraction of the speed of light and eventually accesses all the cosmological resources that are in principle accessible from our starting point. It's a finite quantity because of the positive cosmological constant: looks like we can only access a finite amount of stuff. But once you've started that, once you're an intergalactic empire, it looks like it could just keep going with high probability to this natural vision.

We can define the concept of an existential risk as one that fails to realize the potential for value that you could gain by accessing the cosmological commons, either by going extinct or by maybe accessing all the cosmological commons but then failing to use them for beneficial purposes or something like that.

That suggests this Maxipok principle that Beckstead also mentioned: /Maximize the probability of an OK outcome/. That's clearly, at best, a rule of thumb: it's not meant to be a valid moral principle that's true in all possible situations. It's not that. In fact, if you want to go away from the original principle you started [with] to something practically tractable, you have to make it

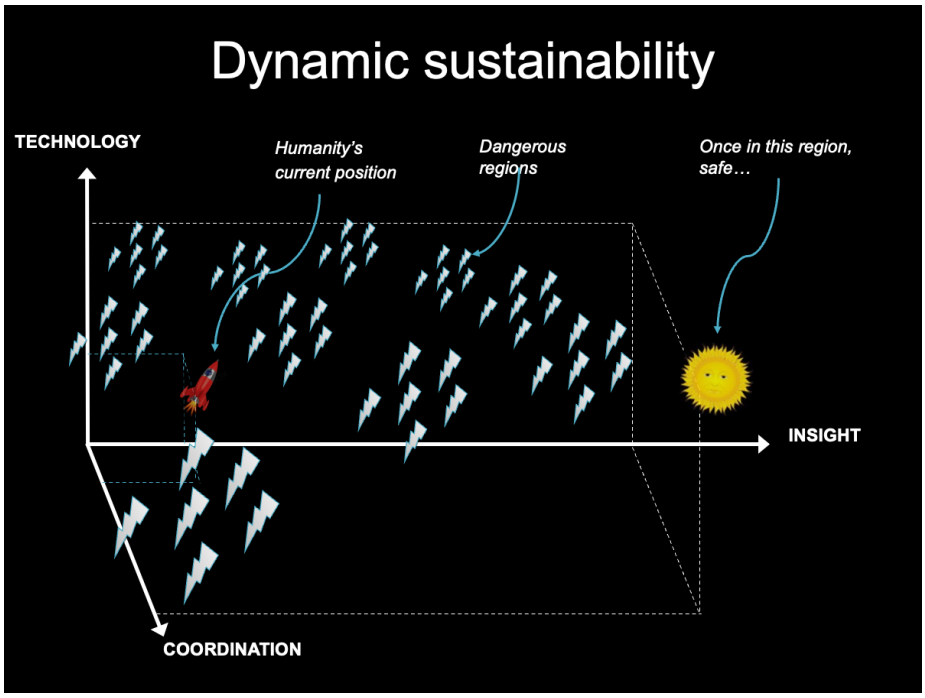
contingent on various empirical assumptions. That's the trade-off there: you want to make as weak assumptions as you can and still move it as far as possible towards being tractable as you can.

I think this is something that makes a reasonable compromise there. In other words, take the action that minimizes the integral of existential risk that humanity will confront. It will not always give you the right answer, but it's a starting point. There are different things to the ones that Beckstead mentioned, there could be other scenarios where this would give the wrong answer: if you thought that there was a big risk of hyper existential catastrophe like some hell scenario, then you might want to increase level of existential risks slightly in order to decrease the risk that there would not just be an existential catastrophe but hyper existential catastrophe. Other things that could come into it are trajectory changes that are less than drastic and just shift slightly.

For present purposes, we could consider the suggestion of using the Maxipok rule as our attempt to define the value function for utilitarian agents. Then the question becomes, If you want to minimize existential risk, what should you do? That is still a very high-level objective. We still need to do more work to break that down into more tangible components.

I'm not sure how well this fits in with the rest of the presentation.

[laughter]




I have this nice slide from another presentation. It's a different way of saying

some of what I just said: instead of thinking about sustainability as is commonly known, as this static concept that has a stable state that we should try to approximate, where we use up no more resources than are regenerated by the natural environment, we need, I think, to think about sustainability in dynamical terms, where instead of reaching a state, we try to enter and stay on a trajectory that is indefinitely sustainable in the sense that we can contain it to travel on that trajectory indefinitely and it leads in a good direction.

An analogy here would be if you have a rocket. One stable state for a rocket is on the launch pad: it can stand there for a long time. Another stable state is if it's up in space, it can continue to travel for an even longer time, perhaps, if it doesn't rust and stuff. But in mid-air, you have this unstable system. I think that's where humanity is now: we're in mid-air. The static sustainability concept suggests that we should reduce our fuel consumption to the minimum that just enables us to hover there. Thus, maybe prolong the duration in which we could stay in our current situation, but what we perhaps instead should do is maximize the fuel consumption so that we have enough thrust to reach escape velocity. (And that's not a literal argument for burning as much fossil fuel as possible. It's just a metaphor.)

[laughter]

The point here is that to have the best possible condition, we need super advanced technology: to be able to access the cosmic commons, to be able to cure all the diseases that plague us, etc. I think to have the best possible world, you'll also need a huge amount of insight and wisdom, and a large amount of coordination so as to avoid using high technology to wage war against one another, and so forth. Ultimately, we would want a state where we have huge quantities of each of these three variables, but that leaves open the question of what we want more from our consideration. It might be, for example, that we would want more coordination and insight before we have more technology of a certain type. So that before we have various powerful technologies, we would first want to make sure that we have enough peace and understanding to not use them for warfare, and that we have enough insight and wisdom not to accidentally blow ourselves up with them.

A superintelligence, clearly, seems to be something you want in utopia—it's a very high level of technology—, but we might want a certain amount of insight before we develop superintelligence, so we can develop it in the correct way.

One can begin to think about, as in analogy with the computer test situation, if there are different features that one could possibly think of as components of this evaluation function for the utilitarian, the Maxipok.

This principle of differential technological development suggests that [we should] retard [the] development of dangerous and harmful technologies—one's that raise existential risk, that is— and accelerate technologies that reduce existential risks.

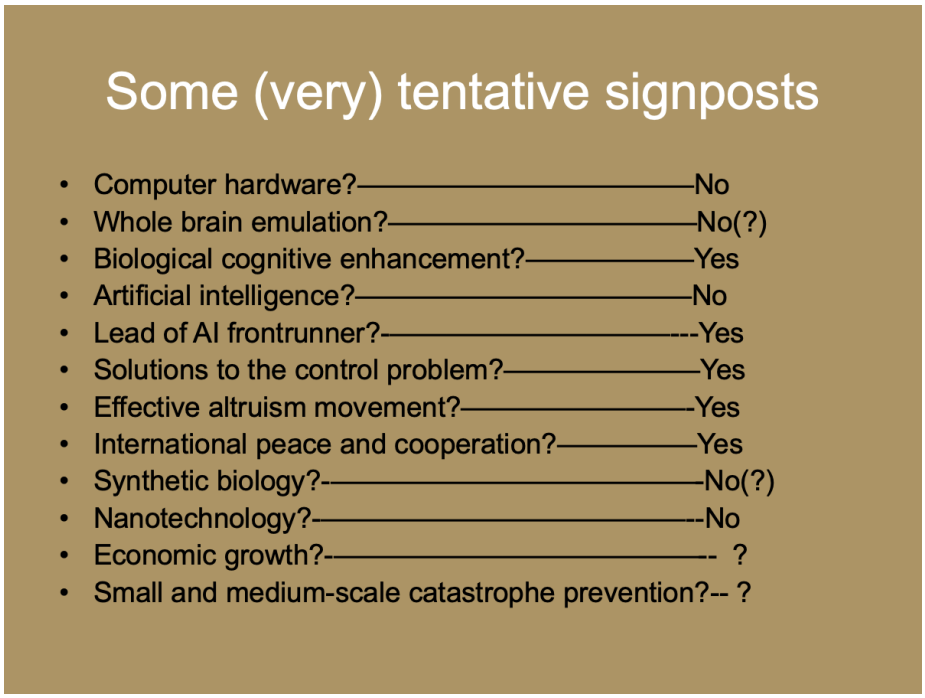
Here is our first sketch, this is not a final answer, but one may think that we want a lot of wisdom, we want a lot of international peace and cooperation, and with regards to technologies, it gets a little bit more complicated: we want faster progress in some technology areas, perhaps, and slower in others. I think those are three broad kinds of things one might want to put into the one's evaluation function.

This suggests that one thing to be thinking about in addition to interventions or causes, is the signature of different kinds of things. An intervention should be sort of high leverage, and a cause area should promise high leverage interventions. It's not enough that something you could do would do good, you also want to think hard about how much good it could do relative to other things you could do. There is no point in thinking about causes without thinking about how do you see all the low hanging fruit that you could access. So a lot of the thinking is about that.

But when we're moving at this more elevated plane, this high altitude where there are these crucial considerations, then it also seems to become valuable to think about determining the sign of different basic parameters, maybe even where we are not sure how we could affect them. (The sign being, basically, Do we want more or less of it?) We might initially bracket questions as to leverage here, because to first orient ourselves in the landscape we might want sort of postpone that question a little bit in this context. But a good signpost—that is a good parameter of which we would like to determine the signature—would have to be visible from afar. That is, if we define some quantity in terms that still make it very difficult for any particular intervention to say whether it contributes positively or negatively to this quantity that we just defined, then it's not so useful as a signpost. So, "maximize expected value", say, is the quantity they could define. It just doesn't help us very much, because whenever you try to do something specific you're still virtually as far away as you had been. On the other hand, if you set some more concrete objective, like maximize the number of people in this room, or something like that, we can now easily tell like how many people there are, [and] we have ideas about how we could maximize it. So any particular action we think of we might easily see how it fares on this objective of maximizing the people in this room. However, we might feel it's very difficult to get strong reasons for knowing whether more people in this room is better, or whether there is some inverse [relationship]. A good signpost would strike a reasonable compromise between being visible from afar and also being such that we can have strong reason to be sure of its sign.

## Some tentative signposts

Here are some very tentative signposts: they're tentative in my own view, and I guess there might also be a lot of disagreement among different people. So these are more like areas for investigation. But it might be useful just to show how one might begin to think about it.




*Do we want faster progress in computer hardware or slower progress?* My best guess there is that we want slower progress. And that has to do with the risks from the machine intelligence transition. Faster computers would make it easier to make AI, which (a) would make them happen sooner probably, which seems perhaps bad in itself because it leaves less time for the relevant kind of preparation, of which there is a great need; and (b) might reduce the skill level that would be required to produce AI: with a ridiculously large amount of computing power you might be able to produce AI without really knowing much about what you're doing; when you are hardware-constrained you might need more insight and understanding, and it's better that AI be created by people who have more insight and understanding.

This is not by any means a knockdown argument, because there are other existential risks. If you thought that we are about to go extinct anytime soon, because somebody will develop nanotechnology, then you might want to sort of try the AI wildcard as soon as possible. But all-things-considered this is my current best guess. These are the kinds of reasoning that one can engage in.

*Whole brain emulation?* We did a long, big analysis of that. More specifically, not whether we want to have whole brain emulation, but whether we want to have more or less funding for whole brain emulation, more or less resources for developing that. This is one possible path towards machine superintelligence, and for complicated reasons, my guess is "No", but that's even more uncertain, and we have a lot of different views in our research group on that. (In the

discussion, if anybody is interested in one particular one, we can zoom in on that.)

*Biological cognitive enhancement of humans?* My best guess there is that we want faster progress in that area.

So with these three—I talk more about them in the book—and this one as well [AI]: I think we want AI probably to happen a little bit slower than it's likely to do by default.

Another question is:

*If there is one company or project or team that will develop the first successful AI, how much ahead does one want that team to be to the second team that is trying to do it?* My best guess is that we wants it to have a lot of lead, many years ideally, to enable them to slow down at the end to implement more safety measures, rather than being in the tight tech race.

*Solutions to the control problem for AI?* I think we want faster progress in that, and that's one of our focus areas, and some of our friends from the Machine Intelligence Research Institute are here, also working hard on that.

*The effective altruism movement?* I think that looks very good in many ways, robustly good, to have faster, better growth in that.

*International peace and cooperation?* Looks good.

[laughter]

*Synthetic biology?* I think it looks bad. We haven't thought as carefully about that, so that could change, but it looks like there could be x-risks from that, although [it may be] also beneficial. Insofar as it might enable improvements in cognitive enhancement, there'll be a kind of difficult trade-off.

*Nanotechnology?* I think it looks bad: [we want] slower progress towards that.

*Economic growth?* Very difficult to tell the sign of that, in my view. And within a community of people have thought hard about that are, again, different guesses as to the sign of that.

*Small and medium scale catastrophe prevention?* Also looks good. So global catastrophic risks falling short of existential risk. Again, very difficult to know the sign of that. Here we are bracketing leverage at all, even just knowing whether we would want more or less, if we could get it for free, it's non-obvious. On the one hand, small-scale catastrophes might create an immune response that makes us better, puts in place better safeguards, and stuff like that, that could protect us from the big stuff. If we're thinking about medium-scale catastrophes that could cause civilizational collapse, large by ordinary standards but only medium-scale in comparison to existential catastrophes, which are large in this context, again, [it is] not totally obvious what the sign of thatis: there's a lot more work to be done to try to figure that out. If recovery looks very likely, you might then have guesses as to whether the recovered civilization would be more

likely to avoid existential catastrophe having gone through this experience or not.

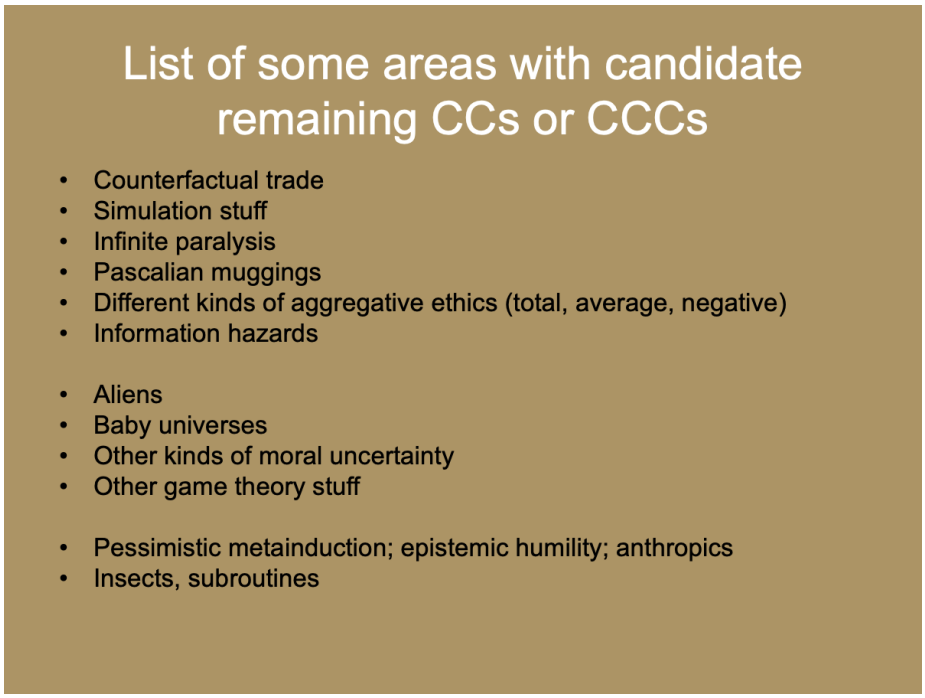
So these are the parameters that one can begin to think about. One doesn't realize just how difficult it is, even some parameters that from an ordinary common-sense point of view seem kind of obvious, actually turn out to be quite non-obvious once you start to think through the way that they're all supposed to fit together.

Suppose you're an administrator here in Oxford, you're working in the Computer Science department, and you're the secretary there. Suppose you find some way to make the department run slightly more efficiently: you create this mailing list so that everybody can, when they have an announcement to make, just email it to the mailing list rather than having to put in each person individually in the address field. And that's a useful thing, that's a great thing: it didn't cost anything, other than one-off cost, and now everybody can go about their business more easily. From this perspective, it's very non-obvious whether that is, in fact, a good thing. It might be contributing to AI—that might be the main effect of this, other than the very small general effect on economic growth. And it might probably be that you have made the world worse in expectation by making this little efficiency improvement. So this project of trying to think through this it's in a sense a little bit like the Nietzschean *Umwertung aller Werte*—the revaluation of all values—project that he never had a chance to complete, because he went mad before.

[laughter]

## Possible areas with additional crucial considerations

So, these are some kinds of areas—I'm not going to go into all of these, I'm just giving examples of the kinds of areas where today it looks like there might still be crucial considerations. This is not an exhaustive list by any means, and we can talk more about some of those. They kind of go from more general and abstract and powerful, to more specific and understandable by ordinary reasoning.




To just pick an example: *insects*. If you are a classical utilitarian, this consideration arises within the more mundane—we're setting aside the cosmological commons and just thinking about here on Earth. If insects are sentient then maybe the amount of sentience in insects is very large because there are so very, very many of them. So that maybe the effect of our policies on insect well-being mind trump the effect of our policies on human well-being or animals in factories and stuff like that. I'm not saying it does, but it's a question that is non-obvious and that could have a big impact.

[Or take another example:] *subroutines*. With certain kinds of machine intelligence there are processes, like reinforcement learning algorithms and other subprocesses within the AI, that could turn out to have moral status in some way. Maybe there will be hugely large numbers of runs of these subprocesses, so that if it turns out that some of these kinds of things count for something, then maybe the numbers again would come to dominate.

## Some partial remedies

Each of these is a whole workshop on its own, so it's not something we can go into. But what can one do if one suspects that there might be these crucial considerations, some of them not yet discovered? I don't have a crisp answer to that. Here are some *prima facie* plausible things one might try to do a little bit of:

- *Don't act precipitously, particularly in ways that are irrevocable.*
- *Invest in more analysis to find and assemble missing crucial considerations.* That's why I'm doing the kind of work that I'm doing, and the rest of us are also involved in that enterprise.
- *Take into account that expected value changes are probably smaller than they appear.* If you are a utilitarian, let's say you think of this new argument that has this radical implication for what you should be doing, the first instinct might be to radically change your expected utility of different practical policies in light of this new insight. But maybe when you reflect on the fact that there are new crucial considerations being discovered every once in awhile, maybe you should still change your expected value, but not as much as it seems you should the first time. You should reflect on this at the meta level.
- *[Take into account fundamental moral uncertainty.]* If we widen our purview to not just consider utilitarianism, as we should consider things from a more general unrestricted normative perspective, then something like the Parliamentary Model for taking normative uncertainty into account looks fairly robust. This is the idea that if you are unsure as to which moral theory is true, then you should assign probabilities to different moral theories and imagine that there were a parliament where each moral theory got to send delegates to that parliament in proportion to their probability. Then in this imaginary parliament, these delegates from the different moral theories discuss and compromise and work out what to do. And then you should do that what that moral parliament of yours would have decided, as a sort of metaphor. The idea is that, other things equal, the more probability a moral theory has, the greater its say in determining your actions, but there might also be these trades between different moral theories which I think Toby talked about in his presentation. This is one metaphor for how to conceive of those traits. It might not be exactly the right way to think about fundamental normative uncertainty, but it seems to be close in many situations, and it seems to be relatively robust in the sense of being unlikely to have a totally crazy implication.
- *Focus more on near-term and convenient objectives.* To the extent that one is despairing about having any coherent view about how to go about maximizing aggregative welfare in this cosmological context, the greater it seems the effective voice of other types of things that one might be placing weight. So if you're partly an egoist and partly an altruist, then if you say that the altruistic component is on this kind of deliberation ladder then maybe you should go more with the egoistic part, until and unless you can find stability in your altruistic deliberations.
- *Focus on developing our capacity as a civilization to wisely deliberate on these types of things.* To build up our capacity, rather than pursuing very

specific goals, and by capacity in this context it looks like perhaps we should focus less on powers and more on the propensity to use powers as well. This is still quite vague, but something in that general direction seems to be robustly desirable. Certainly, you could have a crucial consideration that's turned up to show that that was the wrong thing to do, but it still looks like a reasonable guess.

That's it. Thanks.