Some Key Ways in Which I've Changed My Mind Over the Last Several Years

Summary of this document on the Open Philanthropy Blog

This document discusses three interrelated topics on which I've changed my mind fairly dramatically over the last several years:

1. The importance of potential risks from advanced artificial intelligence, particularly the value alignment problem.
2. The potential of many of the ideas, and people, associated with the effective altruism community.
3. The general properties I should look for when assessing how promising an idea or intervention is - and in particular, the relative importance of metrics and "feedback loops" compared to other properties one might look for.

Earlier this year, I discussed my current views on #1 and #3, without much discussion of how those views differ from views I used to hold. This document is primarily of interest for those who are either (a) interested in a (simplified) summary of how my personal thinking has *changed*; or (b) interested in topic #2, which is deeply intertwined with the other two (hence my choice to cover it in this document).

This is very much a document about my personal thinking and worldview. I think it is appropriate to write about these topics because I believe that philanthropy - especially hits-based philanthropy - is unavoidably driven by a large number of judgment calls, and at the Open Philanthropy Project we've explicitly designed our process so that key individuals (including myself) carry major weight in the strategies we pursue, the causes we prioritize and the grants we ultimately make. (More on this model here, here and here.) At for-profit investing firms, such as hedge funds and VCs, changes in how leadership sees the world can be very important for investment strategy; I think our case is analogous in that respect.

I cover the three topics listed above in increasing order of generality, and it becomes increasingly difficult to identify and discuss all of the things that have gone into my changing views, so I discuss them at decreasing levels of detail.

Evolution of my views on potential risks of advanced artificial intelligence

Initial position

In 2007, the year Elie Hassenfeld and I started GiveWell, we met several people from a community that would later become identified with effective altruism. They introduced us to the idea of potential risks from advanced artificial intelligence, and in particular to the views of Eliezer Yudkowsky and MIRI (then called the Singularity Institute for Artificial Intelligence) regarding the value alignment problem.

My initial reaction to these ideas, as I recall, was as follows:

- I found the topic fascinating and fun to discuss. I accepted that extremely powerful AI systems seemed possible in principle, that it was easy to imagine AI advances being extremely transformative for society (to the point of being among the most important developments in history), and that there was no particular reason to assume that a powerful AI system would necessarily be "friendly" (Eliezer's term) in the sense of having objectives aligned with human values. Because of these points, I could see the basic logic behind why one might see this topic and MIRI's work as important, and I believed that the general issue deserved more serious attention than it was getting.
- There were parts of MIRI's views I did not follow, particularly the seeming assumption that future relevant AI systems would operate as "agents" rather than "tools" (more in my 2012 post on the subject). Until mid-2011 or so, I didn't consider this disagreement important and didn't emphasize it, due to the next bullet point.
- More importantly, I felt unqualified to judge whether MIRI's ideas were important. I felt that **MIRI's lack of impressive endorsements from people with relevant-seeming expertise was the most important data point about it**, and indicated that there were likely strong arguments - even if I did not know what they were - against putting resources into the kinds of concerns it indicated. I interpreted the general degree to which MIRI's views were seen as "wacky" and "silly" to a broad variety of people I spoke with in a similar way: I had a background view that ideas with so little acceptance were probably that way for a reason, and that finding the people most qualified to comment on them would likely result in finding strong counterarguments. (More on my general heuristics around "wacky" ideas in the next section.)
- Reinforcing this view was the impression that MIRI staff and supporters had unrealistic pictures of their abilities, unrealistic plans, and an unwarranted resistance to imposing measurable "tests" on their views. Some reasons I believed this were given in my 2012 post on the subject, though in earlier years, there was another important data point not mentioned in that post: the fact that MIRI staff seemed to consider it a live possibility that they would build artificial general intelligence themselves, achieving this with a substantial lead time on any other group. More broadly, I had a generally skeptical view of the community that introduced me to MIRI, discussed more in the next section.
- It seemed to me that most of the MIRI supporters I came into contact with agreed with

much of the previous point, and based their arguments for supporting MIRI primarily on the argument along the lines of "This goal is so important that even an infinitesimal chance of success would make it the best charity." More on this idea, and why I rejected it, in my 2012 post. For some time, I thought that "how to handle Pascal's Mugging type scenarios" was the main disagreement between me and most MIRI supporters - even though MIRI itself was the source of the "Pascal's Mugging" term, and consistently opposed arguments of this type.

I think the best representation of the views I held between 2007-2011 is the document linked from this 2011 post to GiveWell's (now discontinued) public discussion group. The only public content I'm aware of discussing my views prior to this date is here and here and seems broadly consistent with (though vaguer than) that document.

In 2011, I started to put more effort into thinking about MIRI (though still generally on personal time rather than work time). I'm not entirely sure why this was, but I believe that (a) I was starting to think about how GiveWell might evaluate ideas that couldn't fit in the "proven, cost-effective, scalable" framework we were using (we started thinking about the possibility of GiveWell Labs in February or March of 2011 and announced it in September 2011); (b) The community of people that had introduced me to MIRI seemed to be getting more interested in ideas related to effective altruism (more below) and in GiveWell in particular, and I was getting more questions about how to think about donating to MIRI.

I had a number of discussions about the "tool vs. agent" idea, and I ended up feeling that none of the MIRI supporters I spoke with had good explanations for why we should consider the "agent" framework likely. Furthermore, when talking with friends of mine who were outside the community but worked in technical fields, it seemed they often found the "tool" idea intuitive. I thus came to believe that this distinction was a very strong candidate for "good argument against MIRI's work that explains MIRI's lack of support from people with relevant-seeming expertise."

In 2012, I wrote up my full set of reasons for being skeptical of MIRI at length. These included both detailed arguments (e.g. about tools vs. agents) and MIRI's lack of impressive endorsements and/or achievements. I felt the lack of impressive endorsements and/or achievements was a strong indicator that there was *some* very strong set of arguments against supporting its work, one held by whatever community of computer scientists had the most relevant expertise, and my best guess was that these arguments were related to the tool-agent distinction.

What changed

I think the most important change for me between 2012-2015 was in how I thought about the question, "Why are there no (or few) people with relevant-seeming expertise who endorse MIRI's arguments about the value alignment problem?"

As of 2012, I believe I had a model of the world that was roughly along the lines:

- The more important a topic, the more likely it is that there is some community of people with deep relevant expertise and credentials that can mark this expertise. In other words,

there ought to be "experts" on the topics MIRI was discussing, and MIRI didn't seem to be the experts.
- The lack of endorsements from such experts was most easily explained by the hypothesis that relevant experts had strong counterarguments to MIRI's points, and were simply not bothering to engage.

I believe my take on this topic started to change noticeably when I moved to San Francisco in 2013, and started meeting more people who worked professionally on AI and machine learning and could talk to me about the community around these fields. I started to get the sense that my friends' colleagues in the AI and machine learning fields were largely uninterested in the topics MIRI was interested in (and/or reluctant to talk about them for fear of appearing unserious to others), without necessarily having concrete reasons for dismissing the types of risks MIRI discussed.

In January 2014, we started a shallow investigation of potential risks of advanced artificial intelligence. We talked to a few AI researchers about their views, and in June of that year, I wrote:

> *We are currently working on trying to understand whether the seeming lack of activity comes from a place of 'justified confidence that action is not needed now' or of 'lack of action despite a reasonable possibility that action would be helpful now.' My current guess is that the latter is the case, and if so I hope to make this cause a priority.*

I feel that the most dramatic update on this front came from the publication of Superintelligence and the general reaction to it:

- The book generated a great deal of positive attention, including from public figures.
- At least one mainstream AI researcher, Stuart Russell, specifically agreed that there was a "value alignment problem."
- To the extent I saw explicit criticisms of *Superintelligence*'s ideas, including from AI and machine learning researchers, I didn't feel the criticisms were compelling (more).
- I didn't see any AI or machine learning researchers using arguments along the lines of my "tool vs. agent" distinction, which lowered the likelihood (in my mind) that this distinction was important at all, and especially that it could explain the previous lack of interest in these issues.
- From talking to friends working on AI and machine learning research, I got the impression that many of the points raised in the book were new to their colleagues in those fields.
- All of this surprised me greatly. When I'd skimmed *Superintelligence* (prior to its release), I'd felt that its message was very similar to - though more clearly and carefully stated than - the arguments MIRI had been making without much success. I'd have strongly predicted that the book would be widely ignored and/or panned by people in the AI and machine learning communities.

We paused our investigation of potential risks from advanced artificial intelligence when we learned of an upcoming conference on the subject, which we planned to send representatives to and which we hoped would give us a clearer sense of the state of the arguments. As stated in early 2015, this conference was another major update, solidifying the idea that my earlier view (discussed above) had been wrong. To be clear, I didn't believe the open letter following the conference was a clear *endorsement* of the arguments in *Superintelligence* - the letter was broad in its statements - but I did believe, based on everything I knew including reports from the conference, that my earlier guesses about strong latent counterarguments had been falsified.

This was the most important factor in my changing views over this period, but there were some others as well:

- I learned more about machine learning, particularly deep learning and reinforcement learning, and the general excitement around these areas rose as well. Reinforcement learning seems like a framework in which it may be much easier and faster to build an effective "agent" than to build an effective "tool"; additionally, both deep learning and reinforcement learning seem to make it extremely hard to achieve the kind of algorithmic transparency that I had guessed is generally possible to achieve. I had previously recognized the general possibility of both "algorithmic transparency is hard to achieve" and "agents are easier to build than tools," but intuitively felt both were unlikely; seeing the excitement around deep learning and reinforcement learning changed my mind on this point. (This isn't to say that I now think the "tool" approach has no potential - just that I see a greater risk that this approach will prove infeasible than I used to.)
- My general views about the other two topics covered in this document (below) shifted, and both of these played into how I thought about working on potential risks from advanced AI. As I became more positive on the effective altruism community, I became more inclined to believe that arguments associated with this community had been reasoned out well and carefully thought through; as I became more positive on the potential of high-risk, long-term projects, I saw more potential in trying to anticipate potential risks from advanced AI.
- MIRI improved its operations, addressing some of the criticisms I made of it as an organization (particularly those relating to a seeming lack of focus) and improving my impression of it generally.

Consequences

By early 2015, I felt that the value alignment problem was a real risk, and that I had misjudged what it meant that there was so little interest from people with relevant-seeming expertise. My new interpretation (which I still hold) is that **there simply is no mainstream academic or other field (as of today) that can be considered to be "the locus of relevant expertise" regarding potential risks from advanced AI.** These risks involve a combination of technical and social considerations that don't pertain directly to any recognizable near-term problems in the world, and aren't naturally relevant to any particular branch of computer science. There is no one who I think clearly qualifies as an expert on such matters, and no one I'm aware of who has clearly put more thought into the relevant issues than Nick Bostrom or relevant staff at MIRI.

Learning this was a major update for me. I was very surprised that an issue so potentially important has, to date, commanded so little attention - and that the attention it has received has been significantly (though not exclusively) due to people in the effective altruism community.

At this point, I still was not sure how to think about the overall appeal of the cause, according to Open Philanthropy Project criteria. It seemed possible to me that the reception of *Superintelligence*, and the increasing attention that followed (including FLI's open letter and

Elon Musk's gift), meant that this area was no longer neglected. (It seemed possible that the developments that had convinced me had convinced others as well, to the point where the cause was no longer promising just as I saw the case for it.) Over the course of 2015, we first closely investigated the FLI request for proposals and ultimately decided to fund it, and then contracted with Daniel Dewey to create a landscape of the field as a whole. Over that period, I also significantly changed my mind about how *soon* we should expect transformative AI. These developments raised the appeal of the cause further in my mind, and earlier this year I argued that it represents an outstanding philanthropic opportunity.

Next time a cause this potentially promising emerges, I hope to recognize it faster. With that in mind:

- I'm less likely to interpret "lack of endorsements from people who clearly qualify as experts" as "evidence that an argument is invalid."
- I'm more likely to put weight on arguments that are getting significant attention from the relevant people in the effective altruism community.
- I've altered my general approach to thinking about arguments that sound like they could be very important, but also seem very out of line with conventional wisdom, unsupported by people with clear expertise, and generally "wacky." It's true both now and before that I'm willing to pursue such ideas as long as I've investigated them thoroughly, and that I'm unwilling to bet on any ideas that I haven't investigated much. But going forward, my interest in investigating ideas with these properties - and the time I'm willing to allocate to such investigation - will be higher than previously. And the burden of proof I assume for determining that ideas with these properties are failing to gain acceptance out of lack of engagement (rather than strong counterarguments) will be lower.

Evolution of my views on the effective altruism community

Note: this section focuses on the parts of the effective altruist community that I did *not* initially encounter as people donating to, or spreading the word about, GiveWell and its top charities. For simplicity, I use "effective altruism community" to refer to this set of people, even though there are others (such as Peter Singer) who would not be well described by this section.

Initial position

As mentioned above, in 2007 I started meeting people who would later become associated with the effective altruism community. I'll refer to these people as "pre-effective-altruists," since not only was the term "effective altruism" not in use, but altruism wasn't clearly (to me) a central focus of the community at that time. The writings I most associated with this community were the blog posts on Overcoming Bias and later Less Wrong, which were mostly about rationality. It's worth noting that the people I discuss in this context aren't necessarily representative of the community as a whole; they're representative of my early experiences with the community.

The people I was meeting - such as Michael Vassar and Carl Shulman - had unfamiliar and intriguing ideas about how to do as much good as possible. They pointed me to Nick Bostrom's argument that reducing existential risks[1] was the most promising way to do good, and made the case for potential risks from advanced AI as a particularly important cause toward that end. This was one of the reasons I found the pre-effective-altruist community interesting. Another was that I enjoyed Overcoming Bias, and - relatedly - I found this community's style of communicating very natural and intuitively appealing. I liked the quest to reason as explicitly as possible about why we believed what we believed, which included casting our thinking in terms of theoretically appealing formalisms such as Bayes's rule, doing heavy introspection and writing at length about it, and tending to de-emphasize politeness in favor of intense discussion and debate.

These were the reasons I found the community appealing and spent some personal time getting to know the people in it (mostly those in New York), including attending the 2011 Singularity Summit. At the same time, I strongly doubted that I should ultimately put much weight on this community's ideas, or alter my personal or professional priorities based on them.

The basic reasons for this were, as I recall:

- Primarily, this community wasn't associated with anything that seemed to me to qualify as impressive accomplishments, impressive associations and endorsements, or impressive execution toward any particular goal. Furthermore, I had various experiences that seemed to suggest that people in this community were actively difficult to work with, prone to failing at relatively achievable goals, and worse at running organizations than most people who run organizations. (The best public record of these general sentiments is in the second half of my 2012 post on MIRI, though it doesn't contain all of the examples I have in mind.)
- Despite the above point, the projects people in this community were working on looked extraordinarily ambitious, and fundamentally difficult in the sense of involving very little in the way of feedback loops. My 2011 conversation with MIRI gives some flavor of this issue. I felt that people were systematically overestimating what they could realistically accomplish.
- The arguments I associated with this community tended to be very abstract and theoretical, and to reach very unusual conclusions based on what seemed to be simple logic unaccompanied by deep investigation. A good example of this is Astronomical Waste, an essay of about five pages that concludes, "The utilitarian imperative 'Maximize expected aggregate utility!' can be simplified to the maxim 'Minimize existential risk!'" When I tried out these sorts of arguments with friends, the general theme was that they "seem crazy in a way that's hard to refute explicitly." I saw little evidence that anyone in the pre-effective-altruist community had done deep empirical investigation into the relevant issues, or had done enough work toward their goals to discover much of the subtle/hidden obstacles that theoretical arguments couldn't uncover.
- On the details of the arguments, I mostly found things difficult to assess - for example, I didn't know whether there were any realistic threats of human extinction (this was my interpretation of "existential risk" at the time) and whether there were any effective potential interventions to address such threats that weren't already underway via government and other institutions. The only strong candidate people seemed to offer as

[1] Where "existential risk" is defined broadly as a risk "where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential," though this wasn't salient to me at the time.

a neglected area was potential risks from artificial intelligence, and I found their arguments weak for reasons discussed in the previous section.

- I had the impression that this community was far more willing than I was to "accept extraordinary claims without extraordinary evidence" in some sense. I thought this was a likely candidate for the main explanation of our disagreements, and when I wrote with this community in mind, I tended to focus on this topic (as I did here, here and here).

A common view among my pre-effective-altruist contacts was that I was rejecting their arguments based entirely on an "absurdity heuristic": an intuitive, emotional resistance to arguments that seem "wacky" or well outside of conventional wisdom. This may have been a factor, even a major one, but it's worth noting that I explicitly recognized this risk and reflected intensively on whether I was falling prey to such a thing. So regardless of whatever subconscious role this factor played in my thinking, consciously I was using related but different reasoning: I believed the *combination* of highly unusual conclusions, relatively simple and thin (it seemed to me) argumentation, and a seeming lack of real-world experience or accomplishments all suggested that my pre-effective-altruist contacts were failing to use intellectual habits associated with successfully having an impact and were likely missing subtle counter-considerations that such habits would have caught (even if only implicitly).

To some degree, I felt that my positive interest in this community's arguments might indicate a weakness on my part. I suspected that the community and I were both over-weighting arguments that could be expressed in a particular kind of language. I sought to become more like successful, impactful people I could point to, and less like people who seemed to "speak my language" but accomplish little. I felt that changing in this way would be likely to make me more effective and more rational (in the sense of being able to think of the best path to a given objective), and would be likely to result in my having less interest in the pre-effective-altruist community and its views.

What changed

One of the biggest changes is the one discussed above, regarding potential risks from advanced AI. I went from seeing this as a strange obsession of the community to a case of genuine early insight and impact. I felt the community had identified a potentially enormously important cause and played a major role in this cause's coming to be taken more seriously. This development became - in my view - a genuine and major candidate for a "hit", and an example of an idea initially seeming "wacky" and later coming to seem prescient.

Of course, it is far from a settled case: many questions remain about whether this cause is indeed important and whether today's preparations will look worthwhile in retrospect. But my estimate of the cause's likely importance - and, I believe, conventional wisdom among AI researchers in academia and industry - has changed noticeably.

A number of other things have changed as well:

Effective altruism. People in the pre-effective-altruist community had (since I met them) expressed interest in altruism and had unusual ideas about it (as about most topics), but the association in my mind between this community and altruism became much stronger over time.

- In mid-2010 there was an article making the case for giving generously and effectively on Less Wrong, and in late 2010 Less Wrong held a contest for the best article on efficient charity. I believe this topic was rare on Less Wrong prior to that point.
- In 2012, Giving What We Can and 80,000 Hours both became full-time organizations.
- The term "effective altruism" first became prominently used in 2013.

I went from feeling that this community was interesting but largely irrelevant to my work, to feeling that it contained unusually high concentrations of people interested in GiveWell and related issues, to feeling that it contains the closest thing GiveWell (and now the Open Philanthropy Project) has to a natural "peer group" - a set of people who consistently share our basic goal (doing as much good as possible), and therefore have the potential to help with that goal in a wide variety of ways, including both collaboration and critique. I also value other sorts of collaborations and critiques - such as those that question the entire premise of doing as much good as possible, and can bring insights and abilities that we lack - but people who share our basic premises have a unique sort of usefulness as both collaborators and critics, and I've come to feel that the effective altruism community is the most logical place to find them.

Convergence on cause selection. Between 2013-2015, the Open Philanthropy Project applied our criteria of importance, neglectedness and tractability in order to identify focus areas. We've ultimately ended up feeling that the preponderance of the causes that we've seen the most excitement about in the effective altruism community are, in fact, outstanding by these criteria - particularly farm animal welfare and biosecurity and pandemic preparedness (in addition to potential risks from advanced artificial intelligence). They aren't the *only* outstanding causes we've identified, but overall, I've increased my estimate of how well excitement in this community predicts what I will find promising after more investigation.

More interaction with people in the community. We moved to San Francisco in 2013, and I visited the UK in 2014 and 2015, so I've had more opportunities to meet and talk with people in the community than previously. Doing so has caused me to update on a couple of fronts:

- I feel that many people in this community - in particular, key staff at MIRI and FHI - do not make much, if any, recourse to what I think of "Pascal's Mugging" type arguments; in other words, they don't tend to argue things like "I have no reason to think X is possible, but I can't be *sure* it *isn't*, and an infinitesimal probability is enough." They generally expect to be able to give positive arguments for relatively high probability (though this might sometimes mean "at least 1%" or "at least 10%") of key scenarios. (Here's one conversation that caused me to update on this front.)
- Rather than basing their views on simple, thin argumentation and problematic approaches to uncertainty, I believe that at least some of these people have put a great deal of thought and informal investigation into just about all aspects of the issues they focus on - practical, empirical, and philosophical. I've been particularly impressed with Carl Shulman's reasoning: it seems to me that he is not only broadly knowledgeable, but approaches the literature that influences him with a critical perspective similar to GiveWell's.

I anticipate that some will wonder about whether a changing social circle has been a factor as well. I think it is one, though a relatively minor one. I am housemates with Nick Beckstead (formerly with FHI, now with the Open Philanthropy Project), who has much stronger connections to others in the community than I do; some of my increased exposure to various people's views has come via Nick. My other housemates have tended, like me, to be more in the camp of "sympathetic to the goals of effective altruists but not tightly integrated into the community." I've had little social interaction with other people in the community, though a fair amount of professional interaction due to the number of GiveWell and Open Philanthropy Project employees who identify as part of this community.

Accomplishments. Over the last five years or so, it tentatively seems to me that people in the effective altruism community have accomplished or will accomplish more than I expected. I count bringing attention to potential risks from advanced AI as one accomplishment. I feel that several organizations founded in this community have lasted longer and seem better-run than I would have predicted based on my previous views (this includes MIRI, which I believe has improved significantly since 2012). It also seems to me that they have done a good job galvanizing interest in effective altruism and growing the community, which has had effects I've seen directly (more attention, donors, and strong employee candidates for GiveWell and the Open Philanthropy Project). I caveat all this by noting that it seems too early to assess the long-run impact of most things people in this community are trying to do, and I think the update for me on this front has been moderate rather than drastic.

At the same time, I've moved away from my earlier position (described at the end of the previous section) that I ought to be trying to become more similar to people with strong objective accomplishments. This position would have predicted that as I learned more details and did more investigation on topics of interest for effective altruism, I would move toward conventional wisdom (particularly among those with strong objective accomplishments) rather than toward the views of the effective altruism community; the opposite has happened:

- I've spent a good deal more time both with people in the effective altruism community and with various "conventionally successful" people, and I haven't had the feeling that the latter's achievements and ideas for how to create change seem more relevant for the goals we're focused on than those arising from the EA community.
- Having gotten to know both the topics and people's reactions to them better, it now seems to me that people with strong objective accomplishments are often too busy to think carefully about topics of interest to the effective altruism community (which rarely overlap heavily with the work they've been successful in), and/or have optimized their thinking habits for their core work rather than for these topics. For example, successful machine learning researchers have not necessarily had the time or inclination to reflect on the long-term potential risks of advanced AI.
- On topics like this, I think that putting in a lot of thought (as key people in the effective altruism community have) is very important, and that whatever good mental habits accompany having objective accomplishments can't outweigh this.
- My changing views of "feedback loops" (discussed in the next section) play into this change as well.

Changing views on the general properties of promising ideas and interventions, discussed in the next section.

Consequences

When investigating a topic, it's extremely valuable to find someone who has thought about the topic longer and more intensely (and at least equally intelligently and rationally) compared to oneself, with a similar underlying value set driving their views. This can be a huge time-saver in getting started, and seems important if one wants useful critical discussion. Finding such people can be challenging in many cases for the Open Philanthropy Project, since many of the topics we care most about do not have easily identifiable "experts." I now believe that the effective altruism community is the most likely place to find such people for unusual topics that are highly relevant to accomplishing as much good as possible (including questions about the likely moral value of the far future, about the importance/neglectedness/tractability of non-mainstream but high-potential causes, etc.)

I also feel that one of the best things a philanthropist can do is open-endedly support people, and promote the growth of communities, that deeply share the philanthropist's values and goals for the world. (See The Rise of the Conservative Legal Movement for one defense and illustration of this idea; see our thoughts on the Sandler Foundation for another.) I believe the effective altruism community is the most likely place to look for this sort of opportunity in our case. To this end, we are now interested in making grants to support the growth of this community; our work on this front is led by Nick Beckstead. This is a change from our previous position; another factor in the change is that while the likely time cost was a major concern previously, I've now become convinced Nick can do this work with relatively low time investment.

In these two ways, I consider the effective altruism community to be important. That's not to say that I support it unreservedly; I have concerns and objections regarding many ideas associated with it and some of the specific people and organizations within it. I also think the jury is very much out on how much effective altruism organizations have accomplished and will accomplish. But I've changed my mind significantly away from my early impressions that there was not much to learn from people in this community.

Evolution of my views on the general properties of promising ideas and interventions

Initial position

In 2007, I believe I had a general worldview that had strong components of the following ideas:

- Any given person's intellectual reasoning is extremely unreliable and unlikely to reach correct conclusions by itself.
- A much more effective way to arrive at effective ideas and interventions is via "feedback loops": trying something, seeing how it goes, making small adjustments, and trying again many times.
- Someone experienced with feedback loops is likely to have many helpful but subtle and hard-to-formalize intuitions - not just about the specific work they've done, but about how to have impact in general.
- Knowledge and insight tend to be broadly distributed. When seeking feedback on an idea, it is best to hear from a large set of people, since there are many people who might

have something useful to contribute and none who are likely to have a comprehensive view of relevant considerations. Ideas can be refined and improved via feedback loops if getting lots of feedback from lots of directions.

- Society as a whole has an additional important mechanism for generating effective ideas and interventions: many people try different ideas and interventions, and the ones that are successful in some tangible way (e.g. generating a profit) become more prominent, powerful and imitated.
- Some people, organizations and ideas have enormously outsized impact ("hits"). But these "hits" are arrived at through the above processes. When predicting whether a project can have outsized impact, it's very hard and unlikely to be helpful to project the expected effects of the project (for example in lives saved), due to the high uncertainty; it's much more important to consider the people involved and the "feedback loops" that will likely affect them, to determine whether the project will give itself many or few chances to hit upon some unanticipated insight.

Note that I still hold much of this worldview to a significant degree, though less than previously (as I will discuss below).

This general cluster of views was a factor in all of the following:

- I felt that nonprofits had very poor "feedback loops" by default, and that improving them could be an enormously promising path to impact. My hope for GiveWell, and a key reason I was so excited about it, was that it sought to do this by emphasizing transparency, quantification, skepticism and rigorous evaluation. (That said, I expected to discover many problems with the original vision of GiveWell and iterate many times before hitting on something better, and this did happen to some degree; some of our major changes of approach are discussed here and here.) GiveWell's own commitment to transparency was intended partly to help us improve via continuous feedback from many directions, and partly to set a model for other nonprofits to do so.
- It seemed to me that the people in the pre-effective-altruist community were relying primarily on intellectual reasoning, unaccompanied by the prospect of (or experience with) helpful feedback loops, and without much engagement or feedback from the wider world as a whole, particularly people with relevant expertise. The case of potential risks from advanced AI seemed like a particularly strong example of this issue. I believed that these qualities were rarely associated with having much impact.
- I suspected that most nonprofits suffered from somewhat similar issues, as their primary feedback loops seemed essentially unconnected to having the kind of impact they claimed. I suspected that ambitious foundations were likely ineffective for similar reasons.

What changed

Of the three topics discussed here, this is the hardest one to trace the evolution of my thinking on. The set of views described above is applicable across many domains, and thus affected by many kinds of evidence. But here are some major factors I can recall, presented with eye toward conceptual flow rather than in order of importance:

Learning about the history of philanthropy. One of our earliest activities for GiveWell Labs (later the Open Philanthropy Project) was to try to learn about the history of philanthropy; our earliest output, from a scan of 100 claimed successes, is here. Concurrently, we were talking to other major foundations about their activities and past successes and disappointments. Overall, I felt surprised by the number and impressiveness of success stories I saw. Many seemed relatively light on the kind of "feedback loops" I would have expected, and to involve successes of a type that were fairly close to the original vision. And I saw enough "hits" from the very few largest, most prominent foundations that I doubted a pure "survivorship bias" story, i.e., I don't have the sense that these "hits" were merely the lucky few from an enormous sample of different foundations trying different approaches.

Learning more about other relevant history. Particularly over the last couple of years, I've been attempting (mostly on personal time) to become better informed about the key events and drivers behind historical improvements in human empowerment and welfare, including scientific advancement. While I haven't kept good records of this, I've been generally surprised by:

- The fact that a number (though not by any means the majority) of important-seeming breakthroughs seem *not* to have been generated by the dynamics laid out above (not via competition and selection, nor via rapid iteration). I believe that deliberate, strategic government policy has played an important role in many countries' industrialization.[2] Bell Labs, Xerox PARC and DARPA all seem like examples of forward-looking institutions with very little in the way of competition or accountability, which produced great value largely by explicitly aiming for long-run impact via high-risk investments.
- The seemingly strong track record of intellectuals who explicitly based their arguments primarily on (often simple) logical reasoning, empiricism, and belief in natural laws.
  - A particularly strong example is Jeremy Bentham, who pioneered utilitarianism. He is known for using relatively simple logic to reach non-obvious conclusions about morality that many find (and found) repugnant, and I would have expected that this approach would generate many views that look extremely misguided in retrospect. In fact, I haven't come across obvious cases of such views - quite the opposite, Jeremy Bentham seems to have taken a remarkable number of positions that look extremely prescient in retrospect.[3] A generally similar pattern applies to many of the well-known thinkers of the Enlightenment. While many had some beliefs that look misguided in retrospect, the number of such misguided beliefs seems relatively low given how long ago they lived, and overall it seems that they were remarkably prescient for their time.
  - When reading about the history of science,[4] I've had the impression that many

[2] Two books that have influenced my thinking on this subject: Global Economic History: A Very Short Introduction, The Unbound Prometheus

[3] Wikipedia: "He advocated individual and economic freedom, the separation of church and state, freedom of expression, equal rights for women, the right to divorce, and the decriminalising of homosexual acts. He called for the abolition of slavery, the abolition of the death penalty, and the abolition of physical punishment, including that of children. He has also become known in recent years as an early advocate of animal rights."

[4] The book that I've found most useful for this topic is Asimov's Chronology of Science & Discovery.

scientists had remarkably prescient views, seemingly based primarily on intellectual reasoning about fairly sparse facts rather than based on incontrovertible evidence or strong feedback loops. Many such scientists made many substantial contributions, sometimes in disparate fields,[5] which I think somewhat reduces the risk that this is simply a "survivorship bias" phenomenon. I've occasionally looked up famous scientists looking for cases where they were badly wrong on subjects outside the subjects they gained prominence for, and while I have certainly found such cases, I've generally found fewer than I expected to.

    - I also believe that past "futurists" - people who made long-term predictions - have achieved some degree of accuracy. While far from being overwhelmingly or even mostly correct, they haven't been so unreliable and misguided as to make me think that forecasting the long-term future is futile. We hope to publish more on this topic at a later date.

I don't want to overstate the case here - there are many stories of intellectual planning and forecasting gone wrong, and an overwhelming number of contributions to the world that I think are best attributed to feedback loops, competition and selection, etc. I started with a fairly extreme stance that these things (feedback loops, etc) are *necessary* for positive impact, and while I still believe they are very helpful and important, I now believe that there is *some* scope for more speculative, intellectual approaches.

Other factors.

- Through working on GiveWell, I came to see the problem of measurement in nonprofit work as far more difficult than I had anticipated, which should imply - according to the worldview sketched out above - that nonprofits were even less effective than I had anticipated. Yet I updated in the opposite direction. On site visits, I saw many problems and concerns, but fewer than I initially expected to (including when noticing and asking about the work of nonprofits that I wasn't explicitly visiting, and that don't meet GiveWell's criteria). Through our work on the Open Philanthropy Project, I've come to see the case for a variety of nonprofit activities that are hard to measure with the same conclusiveness that we've sought for GiveWell, and where feedback is often more incremental and informal - for example, attracting hires with strong resumes and good reputations, getting attention from media and important figures, etc. I have become less pessimistic about the effectiveness of nonprofits overall, compared to the fairly extreme position I held before. (That said, I think there is still a strong case for GiveWell's top charities' being better bets than the vast majority of alternatives, especially for people who lack the time/capacity/context to do more incremental, informal evaluations of charities. And I think there are still strong reasons to expect disappointing impact for many charities with difficult activities and sparse measurement.)
- I've become more interested in arguments with the general profile of "simple, logical argument with no clear flaws; has surprising and unusual implications; produces reflexive dissent and discomfort in many people." I previously was very suspicious of

[5] For examples of the latter, see Immanuel Kant's work on astronomy or Svante Arrheinus's work on the greenhouse effect.

arguments like this, and expected them not to hold up on investigation, for reasons that I've outlined above, and because of my feeling (discussed in this section) that an argument seeming logical is not much of a point in its favor. However, I now think that arguments of this form are generally worth paying serious attention to until and unless flaws are uncovered, because they often represent positive innovations. Arguments I've seen of this form include the historical case of Bentham's and other Enlightenment thinkers' views, discussed above; the case for potential risks of advanced AI, discussed further above; and the case for several of our other focus areas (which we've investigated thoroughly), such as immigration policy and land use reform.

Consequences

I used to think we should be pessimistic about any intervention or idea that doesn't involve helpful feedback loops and/or isn't the product of useful selective processes. I still think these things (feedback loops, selective processes) are very powerful and desirable; that we should be more careful about interventions that don't involve them; that there is a strong case for preferring charities (such as GiveWell's top charities) that are relatively stronger in terms of these properties; and that much of the effective altruism community, including the people I've been most impressed by, continues to underweight these considerations. However, I have moderated significantly in my view, and I now see a reasonable degree of hope for having strong positive impact while lacking these things, particularly when using logical, empirical, and scientific reasoning. This moderation has made me less resistant to the arguments for working on potential risks of advanced AI, and more broadly to many of the arguments and views advanced by people in the effective altruism community.

I still think it is unwise to bet on ideas and interventions *just because* they seem logical and have no obvious flaws. I generally believe in vetting and investigating arguments thoroughly before taking serious action based on them - something that I think some, but not all, people in the effective altruism community tend to do. For people who have very little time to think about how to do as much good as possible, I think it is highly defensible to take a harder-line stance along the lines of the views expressed at the beginning of this section, and to e.g. support GiveWell's top charities rather than more ambitious and speculative work. I reject arguments along the lines of "Donating to top charities can't be rational if you care about future generations," and still endorse most of what I said along these lines in a 2014 conversation on the subject.

To what extent was I mistaken?

This document has discussed multiple ways in which I believe my previous views were *wrong*, but it hasn't commented much on the extent to which they were *mistaken*. By "mistaken," I mean roughly that I should have, and easily could have, changed my mind faster than I did, and/or that I held to the views I had for reasons such as:

- Non-truth-seeking motives, such as wanting to avoid affiliating with low-status people and/or arguments, or simply being stubborn.
- Failing to understand the basic logic of the views I was rejecting.
- Committing logical fallacies and/or holding inconsistent views such that noticing the inconsistency would have changed my mind near-instantly.

I don’t know of a good way to assess this; subjectively, it doesn't feel like this was the case. When engaging with the ideas discussed in this document, I generally had extensive discussions with their defenders, heard and understood the basic logic, had potential logical fallacies pointed out to me, and quickly and directly considered e.g. the possibility that I was driven by non-truth-seeking motives. Consciously, the views I held were fairly coherent, and fit into a fairly internally consistent worldview. Changing my mind on these topics has, by and large, not consisted of dramatic realizations of concrete logical fallacies or of self-deceiving motives; instead, it has been driven by gradual interrelated evolutions, which in turn have seemed to me to have been driven primarily by new observations. That doesn't mean there was no way in which my views were illogical or muddled, just that it took the weight of new evidence to get me to update. And it doesn't mean I think my views were the best possible with the information available. I think there are many people who had educated themselves better than I with comparable time and effort; knew things and understood topics I did not; and rationally disagreed with me.

I'd add that I think the worldview presented above is a fairly widespread and understandable one. The reasons I've updated are difficult to summarize and far from conclusive. I would not be surprised if there are readers of this document who are coming in with a similar worldview and don't feel inclined to update, and I would not necessarily consider such people unreasonable. After all, it's still very possible that my current views are as flawed as (or more flawed than) my previous ones.

Conclusion

Over the last several years, I have become more positive on the cause of potential risks from advanced AI, on the effective altruism community, and on the general prospects for changing the world through relatively speculative, long-term projects grounded largely in intellectual reasoning (sometimes including reasoning that leads to "wacky" ideas) rather than direct feedback mechanisms. These changes in my thinking have been driven by a number of factors, including by each other.

These changes are a major factor in the fact that the Open Philanthropy Project now takes on work (including supporting the effective altruism community) that I would have been much more skeptical of previously. As discussed at the top of this document, I believe that sort of relationship between personal views and institutional priorities is appropriate given the work we're doing.

I'm not certain that I've been correct to change my mind in the ways described here, and I still have a good deal of sympathy for people whose current views are closer to my former ones, but hopefully this document gives a sense of where the changes have come from.