Best of LessWrong: April 2017

1. Effective altruism is self-recommending
2. Project Hufflepuff: Planting the Flag

Best of LessWrong: April 2017

1. Effective altruism is self-recommending
2. Project Hufflepuff: Planting the Flag

Effective altruism is self-recommending

A parent I know reports (some details anonymized):

> Recently we bought my 3-year-old daughter a "behavior chart," in which she can earn stickers for achievements like not throwing tantrums, eating fruits and vegetables, and going to sleep on time. We successfully impressed on her that a major goal each day was to earn as many stickers as possible.
>
> This morning, though, I found her just plastering her entire behavior chart with stickers. She genuinely seemed to think I'd be proud of how many stickers she now had.

The Effective Altruism movement has now entered this extremely cute stage of cognitive development. EA is more than three years old, but institutions age differently than individuals.

What is a confidence game?

In 2009, investment manager and con artist Bernie Madoff pled guilty to running a massive fraud, with $50 billion in fake return on investment, having outright embezzled around $18 billion out of the $36 billion investors put into the fund. Only a couple of years earlier, when my grandfather was still alive, I remember him telling me about how Madoff was a genius, getting his investors a consistent high return, and about how he wished he could be in on it, but Madoff wasn't accepting additional investors.

What Madoff was running was a classic Ponzi scheme. Investors gave him money, and he told them that he'd gotten them an exceptionally high return on investment, when in fact he had not. But because he promised to be able to do it again, his investors mostly *reinvested* their money, and more people were excited about getting in on the deal. There was more than enough money to cover the few people who wanted to take money *out* of this amazing opportunity.

Ponzi schemes, pyramid schemes, and speculative bubbles are all situations in investors' expected profits are paid out from the money paid in by new investors, instead of any independently profitable venture. Ponzi schemes are centrally managed – the person running the scheme represents it to investors as legitimate, and takes responsibility for finding new investors and paying off old ones. In pyramid schemes such as multi-level-marketing and chain letters, each generation of investor recruits new investors and profits from them. In speculative bubbles, there is no formal structure propping up the scheme, only a common, mutually reinforcing set of expectations among speculators driving up the price of something that was already for sale.

The general situation in which someone sets themself up as the repository of others' confidence, and uses this as leverage to acquire increasing investment, can be called a ***confidence game***.

Some of the most iconic Ponzi schemes blew up quickly because they promised wildly unrealistic growth rates. This had three undesirable effects for the people running the schemes. First, it attracted too much attention – too many people wanted into the scheme too quickly, so they rapidly exhausted sources of new capital. Second, because their rates of return were implausibly high, they made themselves targets for scrutiny. Third, the extremely high rates of return themselves caused their promises to quickly outpace what they could plausibly return to even a small share of their investor victims.

Madoff was careful to avoid all these problems, which is why his scheme lasted for nearly half a century. He only promised *plausibly* high returns (around 10% annually) for a successful hedge fund, especially if it was illegally engaged in insider trading, rather than the sort of implausibly high returns typical of more blatant Ponzi schemes. (Charles Ponzi promised to double investors' money in 90 days.) Madoff showed reluctance to accept new clients, like any other fund manager who doesn't want to get too big for their trading strategy.

He didn't plaster stickers all over his behavior chart – he put a *reasonable* number of stickers on it. He played a ***long game***.

Not all confidence games are inherently bad. For instance, the US national pension system, Social Security, operates as a kind of Ponzi scheme, it is not obviously unsustainable, and many people continue to be glad that it exists. Nominally, when people pay Social Security taxes, the money is invested in the social security trust fund, which holds interest-bearing financial assets that will be used to pay out benefits in their old age. In this respect it looks like an ordinary pension fund.

However, the financial assets are US Treasury bonds. There is no independently profitable venture. The Federal Government of the United States of America is quite literally writing an IOU to itself, and then spending the money on current expenditures, including paying out current Social Security benefits.

The Federal Government, of course, can write as large an IOU to itself as it wants. It could make *all* tax revenues part of the Social Security program. It could issue new Treasury bonds and gift them to Social Security. None of this would increase its ability to pay out Social Security benefits. It would be an empty exercise in putting stickers on its own chart.

If the Federal government loses the ability to collect enough taxes to pay out social security benefits, there is no additional capacity to pay represented by US Treasury bonds. What we have is an implied promise to pay out future benefits, backed by the expectation that the government will be able to collect taxes in the future, including Social Security taxes.

There's nothing necessarily wrong with this, except that the mechanism by which Social Security is funded is obscured by financial engineering. However, this misdirection should raise at least some doubts as to the underlying sustainability or desirability of the commitment. In fact, this scheme was adopted specifically to give people the impression that they had some sort of property rights over their social Security Pension, in order to make the program politically difficult to eliminate. Once people have "bought in" to a program, they will be reluctant to treat their prior contributions as sunk costs, and willing to invest additional resources to salvage their investment, in ways that may make them increasingly reliant on it.

Not all confidence games are intrinsically bad, but dubious programs benefit the most from being set up as confidence games. More generally, bad programs are the ones that benefit the most from being allowed to fiddle with their own accounting. As Daniel Davies writes, in The D-Squared Digest One Minute MBA - Avoiding Projects Pursued By Morons 101:

> **Good ideas do not need lots of lies told about them in order to gain public acceptance.** I was first made aware of this during an accounting class. We were discussing the subject of accounting for stock options at technology companies. […] One side (mainly technology companies and their lobbyists) held that stock option grants should not be treated as an expense on public policy grounds; treating them as an expense would discourage companies from granting them, and stock options were a vital compensation tool that incentivised performance, rewarded dynamism and innovation and created vast amounts of value for America and the world. The other side (mainly people like Warren Buffet) held that stock options looked awfully like a massive blag carried out my management at the expense of shareholders, and that the proper place to record such blags was the P&L account.
>
> Our lecturer, in summing up the debate, made the not unreasonable point that if stock options really were a fantastic tool which unleashed the creative power in every employee, everyone would *want* to expense as many of them as possible, the better to boast about how innovative, empowered and fantastic they were. Since the tech companies' point of view appeared to be that if they were ever forced to account honestly for their option grants, they would quickly stop making them, this offered decent *prima facie* evidence that they weren't, really, all that fantastic.

However, I want to generalize the concept of confidence games from the domain of financial currency, to the domain of social credit more generally (of which money is a particular form that our society commonly uses), and in particular I want to talk about confidence games in the currency of credit for achievement.

If I were applying for a very important job with great responsibilities, such as President of the United States, CEO of a top corporation, or head or board member of a major AI research institution, I could be expected to have some relevant prior experience. For instance, I might have had some success managing a similar, smaller institution, or serving the same institution in a lesser capacity. More generally, when I make a bid for control over something, I am implicitly claiming that I have enough social credit – enough of a track record – that I can be expected to do good things with that control.

In general, if someone has done a lot, we should expect to see an iceberg pattern where a small easily-visible part suggests a lot of solid but harder-to-verify substance under the surface. One might be tempted to make a habit of imputing a much larger iceberg from the combination of a small floaty bit, and promises. But, a small easily-visible part with claims of a lot of harder-to-see substance is easy to mimic without actually doing the work. As Davies continues:

> **The Vital Importance of Audit**. Emphasised over and over again. Brealey and Myers has a section on this, in which they remind callow students that like backing-up one's computer files, this is a lesson that everyone seems to have to learn the hard way. Basically, it's been shown time and again and again; companies which do not audit completed projects in order to see how accurate the original projections were, tend to get exactly the forecasts and projects that

> they deserve. Companies which have a culture where there are no consequences for making dishonest forecasts, get the projects they deserve. Companies which allocate blank cheques to management teams with a proven record of failure and mendacity, get what they deserve.

If you can independently put stickers on your own chart, then your chart is no longer reliably tracking something externally verified. If forecasts are not checked and tracked, or forecasters are not consequently held accountable for their forecasts, then there is no reason to believe that assessments of future, ongoing, or past programs are accurate. Adopting a wait-and-see attitude, insisting on audits for actual results (not just predictions) before investing more, will definitely slow down funding for good programs. But without it, most of your funding will go to worthless ones.

Open Philanthropy, OpenAI, and closed validation loops

The Open Philanthropy Project recently announced a $30 million grant to the $1 billion nonprofit AI research organization OpenAI. This is the largest single grant it has ever made. The main point of the grant is to buy influence over OpenAI's future priorities; Holden Karnofsky, Executive Director of the Open Philanthropy Project, is getting a seat on OpenAI's board as part of the deal. This marks the second major shift in focus for the Open Philanthropy Project.

The first shift (back when it was just called GiveWell) was from trying to find the best already-existing programs to fund ("passive funding") to envisioning new programs and working with grantees to make them reality ("active funding"). The new shift is from funding specific programs at all, to trying to take control of programs without any specific plan.

To justify the passive funding stage, all you have to believe is that you can know better than other donors, among existing charities. For active funding, you have to believe that you're smart enough to evaluate potential programs, just like a charity founder might, and pick ones that will outperform. But buying control implies that you think you're so much better, that even before you've evaluated any programs, if someone's doing something big, you ought to have a say.

When GiveWell moved from a passive to an active funding strategy, it was relying on the moral credit it had earned for its extensive and well-regarded charity evaluations. The thing that was particularly exciting about GiveWell was that they focused on outcomes and efficiency. They didn't just focus on the size or intensity of the problem a charity was addressing. They didn't just look at financial details like overhead ratios. They asked the question a consequentialist cares about: for a given expenditure of money, how much will this charity be able to *improve outcomes*?

However, when GiveWell tracks its impact, it does not track objective outcomes at all. It tracks inputs: attention received (in the form of visits to its website) and money moved on the basis of its recommendations. In other words, its estimate of its own impact is based on the level of trust people have placed in it.

So, as GiveWell built out the Open Philanthropy Project, its story was: We promised to do something great. As a result, we were entrusted with a fair amount of attention and money. Therefore, we should be given more responsibility. We *represented* our

behavior as praiseworthy, and as a result people put stickers on our chart. For this reason, we should be advanced stickers against future days of praiseworthy behavior.

Then, as the Open Philanthropy Project explored active funding in more areas, its estimate of its own effectiveness grew. After all, it was funding more speculative, hard-to-measure programs, but a multi-billion-dollar donor, which was largely relying on the Open Philanthropy Project's opinions to assess efficacy (including its own efficacy), continued to trust it.

What is missing here is any objective track record of benefits. What this looks like to me, is a long sort of confidence game – or, using less morally loaded language, a venture with structural reliance on increasing amounts of leverage – in the currency of moral credit.

Version 0: GiveWell and passive funding

First, there was GiveWell. GiveWell's purpose was to find and vet evidence-backed charities. However, it recognized that charities know their own business best. It wasn't trying to do better than the charities; it was trying to do better than the typical charity donor, by being more discerning.

GiveWell's thinking from this phase is exemplified by co-founder Elie Hassenfeld's Six tips for giving like a pro:

> **When you give, give cash – no strings attached.** You're just a part-time donor, but the charity you're supporting does this full-time and staff there probably know a lot more about how to do their job than you do. If you've found a charity that you feel is excellent – not just acceptable – then it makes sense to trust the charity to make good decisions about how to spend your money.

GiveWell similarly tried to avoid distorting charities' behavior. Its job was only to evaluate, not to interfere. To perceive, not to act. To find the best, and buy more of the same.

How did GiveWell assess its effectiveness in this stage? When GiveWell evaluates charities, it estimates their cost-effectiveness in advance. It assesses the program the charity is running, through experimental evidence of the form of randomized controlled trials. GiveWell also audits the charity to make sure they're actually running the program, and figure out how much it costs as implemented. This is an excellent, evidence-based way to generate a *prediction* of how much good will be done by moving money to the charity.

As far as I can tell, these predictions are untested.

One of GiveWell's early top charities was VillageReach, which helped Mozambique with TB immunization logistics. GiveWell estimated that VillageReach could save a life for $1,000. But this charity is no longer recommended. The public page says:

> VillageReach (www.villagereach.org) was our top-rated organization for 2009, 2010 and much of 2011 and it has received over $2 million due to GiveWell's recommendation. In late 2011, we removed VillageReach from our top-rated list because we felt its project had limited room for more funding. As of November 2012, we believe that that this project may have room for more funding, but we still prefer our current highest-rated charities above it.

GiveWell reanalyzed the data it based its recommendations on, but hasn't published an after-the-fact retrospective of long-run results. I asked GiveWell about this by email. The response was that such an assessment was not prioritized because GiveWell had found implementation problems in VillageReach's scale-up work as well as reasons to doubt its original conclusion about the impact of the pilot program. It's unclear to me whether this has caused GiveWell to evaluate charities differently in the future.

I don't think someone looking at GiveWell's page on VillageReach would be likely to reach the conclusion that GiveWell now believes its original recommendation was likely erroneous. GiveWell's impact page continues to count money moved to VillageReach without any mention of the retracted recommendation. If we assume that the point of tracking money moved is to track the benefit of moving money from worse to better uses, then repudiated programs ought to be counted against the total, as costs, rather than towards it.

GiveWell has recommended the Against Malaria Foundation for the last several years as a top charity. AMF distributes long-lasting insecticide-treated bed nets to prevent mosquitos from transmitting malaria to humans. Its evaluation of AMF does not mention any direct evidence, positive or negative, about what happened to malaria rates in the areas where AMF operated. (There is a discussion of the evidence that the bed nets were in fact delivered and used.) In the supplementary information page, however, we are told:

> Previously, AMF expected to collect data on malaria case rates from the regions in which it funded LLIN distributions: […] In 2016, AMF shared malaria case rate data […] but we have not prioritized analyzing it closely. AMF believes that this data is not high quality enough to reliably indicate actual trends in malaria case rates, so we do not believe that the fact that AMF collects malaria case rate data is a consideration in AMF's favor, and do not plan to continue to track AMF's progress in collecting malaria case rate data.

The data was noisy, so they simply stopped checking whether AMF's bed net distributions do anything about malaria.

If we want to know the size of the improvement made by GiveWell in the developing world, we have their *predictions* about cost-effectiveness, an audit trail verifying that work was performed, and their direct measurement of how much money people gave because they trusted GiveWell. The predictions on the final target – improved outcomes – have not been tested.

GiveWell is actually doing unusually well as far as major funders go. It sticks to describing things it's actually responsible for. By contrast, the Gates Foundation, in a report to Warren Buffet claiming to describe its impact, simply described overall improvement in the developing world, a very small rhetorical step from claiming credit for 100% of the improvement. GiveWell at least sticks to facts *about GiveWell's own effects*, and this is to its credit. But, it focuses on costs it has been able to impose, not benefits it has been able to create.

The Centre for Effective Altruism's William MacAskill made a related point back in 2012, though he talked about the lack of any sort of formal outside validation or audit, rather than focusing on empirical validation of outcomes:

> As far as I know, GiveWell haven't commissioned a thorough external evaluation of their recommendations. […] This surprises me. Whereas businesses have a

> natural feedback mechanism, namely profit or loss, research often doesn't, hence the need for peer-review within academia. This concern, when it comes to charity-evaluation, is even greater. If GiveWell's analysis and recommendations had major flaws, or were systematically biased in some way, it would be challenging for outsiders to work this out without a thorough independent evaluation. Fortunately, GiveWell has the resources to, for example, employ two top development economists to each do an independent review of their recommendations and the supporting research. This would make their recommendations more robust at a reasonable cost.

GiveWell's page on self-evaluation says that it discontinued external reviews in August 2013. This page links to an explanation of the decision, which concludes:

> We continue to believe that it is important to ensure that our work is subjected to in-depth scrutiny. However, at this time, the scrutiny we're naturally receiving – combined with the high costs and limited capacity for formal external evaluation – make us inclined to postpone major effort on external evaluation for the time being.
>
> That said,
>
> - >If someone volunteered to do (or facilitate) formal external evaluation, we'd welcome this and would be happy to prominently post or link to criticism.
> - We do intend eventually to re-institute formal external evaluation.

Four years later, assessing the credibility of this assurance is left as an exercise for the reader.

Version 1: GiveWell Labs and active funding

Then there was GiveWell Labs, later called the Open Philanthropy Project. It looked into more potential philanthropic causes, where the evidence base might not be as cut-and-dried as that for the GiveWell top charities. One thing they learned was that in many areas, there simply weren't shovel-ready programs ready for funding – a funder has to play a more active role. This shift was described by GiveWell co-founder Holden Karnofsky in his 2013 blog post, Challenges of passive funding:

> By "passive funding," I mean a dynamic in which the funder's role is to review others' proposals/ideas/arguments and pick which to fund, and by "active funding," I mean a dynamic in which the funder's role is to participate in – or lead – the development of a strategy, and find partners to "implement" it. Active funders, in other words, are participating at some level in "management" of partner organizations, whereas passive funders are merely choosing between plans that other nonprofits have already come up with.
>
> My instinct is generally to try the most "passive" approach that's feasible. Broadly speaking, it seems that a good partner organization will generally know their field and environment better than we do and therefore be best positioned to design strategy; in addition, I'd expect a project to go better when its implementer has fully bought into the plan as opposed to carrying out what the funder wants. However, (a) this philosophy seems to contrast heavily with how most existing

> major funders operate; (b) I've seen multiple reasons to believe the "active" approach may have more relative merits than we had originally anticipated. [...]
>
> - In the nonprofit world of today, it seems to us that funder interests are major drivers of which ideas that get proposed and fleshed out, and therefore, as a funder, it's important to express interests rather than trying to be fully "passive."
> - While we still wish to err on the side of being as "passive" as possible, we are recognizing the importance of clearly articulating our values/strategy, and also recognizing that an area can be underfunded even if we can't easily find shovel-ready funding opportunities in it.

GiveWell earned some credibility from its novel, evidence-based outcome-oriented approach to charity evaluation. But this credibility was already – and still is – a sort of loan. We have GiveWell's predictions or promises of cost effectiveness in terms of outcomes, and we have figures for money moved, from which we can infer how much we were promised in improved outcomes. As far as I know, no one's gone back and checked whether those promises turned out to be true.

In the meantime, GiveWell then leveraged this credibility by extending its methods into more speculative domains, where less was checkable, and donors had to put more trust in the subjective judgment of GiveWell analysts. This was called GiveWell Labs. At the time, this sort of compounded leverage may have been sensible, but it's important to track whether a debt has been paid off or merely rolled over.

Version 2: The Open Philanthropy Project and control-seeking

Finally, the Open Philanthropy made its largest-ever single grant to purchase its founder a seat on a major organization's board. This represents a transition from mere active funding to overtly purchasing influence:

> The Open Philanthropy Project awarded a grant of $30 million ($10 million per year for 3 years) in general support to OpenAI. This grant initiates a partnership between the Open Philanthropy Project and OpenAI, in which Holden Karnofsky (Open Philanthropy's Executive Director, "Holden" throughout this page) will join OpenAI's Board of Directors and, jointly with one other Board member, oversee OpenAI's safety and governance work.
>
> We expect the primary benefits of this grant to stem from our partnership with OpenAI, rather than simply from contributing funding toward OpenAI's work. While we would also expect general support for OpenAI to be likely beneficial on its own, the case for this grant hinges on the benefits we anticipate from our partnership, particularly the opportunity to help play a role in OpenAI's approach to safety and governance issues.

Clearly the value proposition is not increasing available funds for OpenAI, if OpenAI's founders' billion-dollar commitment to it is real:

> Sam, Greg, Elon, Reid Hoffman, Jessica Livingston, Peter Thiel, Amazon Web Services (AWS), Infosys, and YC Research are donating to support OpenAI. In total, these funders have committed $1 billion, although we expect to only spend a tiny fraction of this in the next few years.

The Open Philanthropy Project is neither using this money to fund programs that have a track record of working, nor to fund a specific program that it has prior reason to expect will do good. Rather, it is buying control, in the hope that Holden will be able to persuade OpenAI not to destroy the world, because he knows better than OpenAI's founders.

How does the Open Philanthropy Project know that Holden knows better? Well, it's done some active funding of programs it expects to work out. It expects those programs to work out because they were approved by a process similar to the one used by GiveWell to find charities that it expects to save lives.

If you want to acquire control over something, that implies that you think you can manage it more sensibly than whoever is in control already. Thus, buying control is a claim to have superior judgment - not just over others funding things (the original GiveWell pitch), but over those being funded.

In a footnote to the very post announcing the grant, the Open Philanthropy Project notes that it has historically tried to avoid acquiring leverage over organizations it supports, precisely because it's *not* sure it knows better:

> For now, we note that providing a high proportion of an organization's funding may cause it to be dependent on us and accountable primarily to us. This may mean that we come to be seen as more responsible for its actions than we want to be; it can also mean we have to choose between providing bad and possibly distortive guidance/feedback (unbalanced by other stakeholders' guidance/feedback) and leaving the organization with essentially no accountability.

This seems to describe two main problems introduced by becoming a dominant funder:

1. People might accurately attribute causal responsibility for some of the organization's conduct to the Open Philanthropy Project.
2. The Open Philanthropy Project might influence the organization to behave differently than it otherwise would.

The first seems obviously silly. I've been trying to correct the imbalance where Open Phil is criticized mainly when it makes grants, by criticizing it for holding onto too much money.

The second really is a cost as well as a benefit, and the Open Philanthropy Project has been absolutely correct to recognize this. This is the sort of thing GiveWell has consistently gotten right since the beginning and it deserves credit for making this principle clear and – until now – living up to it.

But discomfort with being dominant funders seems inconsistent with buying a board seat to influence OpenAI. If the Open Philanthropy Project thinks that Holden's judgment is good enough that he should be in control, why only here? If he thinks that other Open Philanthropy Project AI safety grantees have good judgment but OpenAI doesn't, why not give them similar amounts of money free of strings to spend at their discretion and see what happens? Why not buy people like Eliezer Yudkowsky, Nick Bostrom, or Stuart Russell a seat on OpenAI's board?

On the other hand, the Open Philanthropy Project is right on the merits here with respect to safe superintelligence development. Openness makes sense for weak AI,

but if you're building true strong AI you want to make sure you're cooperating with all the other teams in a single closed effort. I agree with the Open Philanthropy Project's assessment of the relevant risks. But it's not clear to me how often joining the bad guys to prevent their worst excesses is a good strategy, and it seems like it has to often be a mistake. Still, I'm mindful of heroes like John Rabe, Chiune Sugihara, and Oscar Schindler. And if I think someone has a good idea for improving things, it makes sense to reallocate control from people who have worse ideas, even if there's some potential better allocation.

On the other hand, is Holden Karnofsky the right person to do this? The case is mixed.

He listens to and engages with the arguments from principled advocates for AI safety research, such as Nick Bostrom, Eliezer Yudkowsky, and Stuart Russell. This is a point in his favor. But, I can think of other people who engage with such arguments. For instance, OpenAI founder Elon Musk has publicly praised Bostrom's book *Superintelligence*, and founder Sam Altman has written two blog posts summarizing concerns about AI safety reasonably cogently. Altman even asked Luke Muehlhauser, former executive director of MIRI, for feedback pre-publication. He's met with Nick Bostrom. That suggests a substantial level of direct engagement with the field, although Holden has engaged for a longer time, more extensively, and more directly.

Another point in Holden's favor, from my perspective, is that under his leadership, the Open Philanthropy Project has funded the most serious-seeming programs for both weak and strong AI safety research. But Musk also managed to (indirectly) fund AI safety research at MIRI and by Nick Bostrom personally, via his $10 million FLI grant.

The Open Philanthropy Project also says that it expects to learn a lot about AI research from this, which will help it make better decisions on AI risk in the future and influence the field in the right way. This is reasonable as far as it goes. But remember that the case for positioning the Open Philanthropy Project to do this relies on the assumption that the Open Philanthropy Project will improve matters by becoming a central influencer in this field. This move is consistent with reaching that goal, but it is not independent evidence that the goal is the right one.

Overall, there are good narrow reasons to think that this is a potential improvement over the prior situation around OpenAI – but only a small and ill-defined improvement, at considerable attentional cost, and with the offsetting potential harm of increasing OpenAI's perceived legitimacy as a long-run AI safety organization.

And it's worrying that Open Philanthropy Project's largest grant – not just for AI risk, but *ever* (aside from GiveWell Top Charity funding) – is being made to an organization at which Holden's housemate and future brother-in-law is a leading researcher. The nepotism argument is not my central objection. If I otherwise thought the grant were obviously a good idea, it wouldn't worry me, because it's natural for people with shared values and outlooks to become close nonprofessionally as well. But in the absence of a clear compelling specific case for the grant, it's worrying.

Altogether, I'm not saying this is an unreasonable shift, considered in isolation. I'm not even sure this is a bad thing for the Open Philanthropy Project to be doing – insiders may have information that I don't, and that is difficult to communicate to outsiders. But as outsiders, there comes a point when someone's maxed out their moral credit, and we should wait for results before actively trying to entrust the Open Philanthropy Project and its staff with more responsibility.

EA Funds and self-recommendation

The Centre for Effective Altruism is actively trying to entrust the Open Philanthropy Project and its staff with more responsibility.

The concerns of CEA's CEO William MacAskill about GiveWell have, as far as I can tell, never been addressed, and the underlying issues have only become more acute. But CEA is now working to put more money under the control of Open Philanthropy Project staff, through its new EA Funds product – a way for supporters to delegate giving decisions to expert EA "fund managers" by giving to one of four funds: Global Health and Development, Animal Welfare, Long-Term Future, and Effective Altruism Community.

The Effective Altruism movement began by saying that because very poor people exist, we should reallocate money from ordinary people in the developed world to the global poor. Now the pitch is in effect that because very poor people exist, we should reallocate money from ordinary people in the developed world to the extremely wealthy. This is a strange and surprising place to end up, and it's worth retracing our steps. Again, I find it easiest to think of three stages:

1. Money can go much farther in the developing world. Here, we've found some examples for you. As a result, you can do a huge amount of good by giving away a large share of your income, so you ought to.
2. We've found ways for you to do a huge amount of good by giving away a large share of your income for developing-world interventions, so you ought to trust our recommendations. You ought to give a large share of your income to these weird things our friends are doing that are even better, or join our friends.
3. We've found ways for you to do a huge amount of good by funding weird things our friends are doing, so you ought to trust the people we trust. You ought to give a large share of your income to a multi-billion-dollar foundation that funds such things.

Stage 1: The direct pitch

At first, Giving What We Can (the organization that eventually became CEA) had a simple, easy to understand pitch:

> Giving What We Can is the brainchild of Toby Ord, a philosopher at Balliol College, Oxford. Inspired by the ideas of ethicists Peter Singer and Thomas Pogge, Toby decided in 2009 to commit a large proportion of his income to charities that effectively alleviate poverty in the developing world.
>
> […]
>
> Discovering that many of his friends and colleagues were interested in making a similar pledge, Toby worked with fellow Oxford philosopher Will MacAskill to create an international organization of people who would donate a significant proportion of their income to cost-effective charities.
>
> Giving What We Can launched in November 2009, attracting significant media attention. Within a year, 64 people had joined the society, their pledged donations

amounting to $21 million. Initially run on a volunteer basis, Giving What We Can took on full-time staff in the summer of 2012.

In effect, its argument was: "Look, you can do huge amounts of good by giving to people in the developing world. Here are some examples of charities that do that. It seems like a great idea to give 10% of our income to those charities."

GWWC was a simple product, with a clear, limited scope. Its founders believed that people, including them, ought to do a thing – so they argued directly for that thing, using the arguments that had persuaded them. If it wasn't for you, it was easy to figure that out; but a surprisingly large number of people were persuaded by a simple, direct statement of the argument, took the pledge, and gave a lot of money to charities helping the world's poorest.

Stage 2: Rhetoric and belief diverge

Then, GWWC staff were persuaded you could do even more good with your money in areas other than developing-world charity, such as existential risk mitigation. Encouraging donations and work in these areas became part of the broader Effective Altruism movement, and GWWC's umbrella organization was named the Centre for Effective Altruism. So far, so good.

But this left Effective Altruism in an awkward position; while leadership often personally believe the most effective way to do good is far-future stuff or similarly weird-sounding things, many people who can see the merits of the developing-world charity argument reject the argument that because the vast majority of people live in the far future, even a very small improvement in humanity's long-run prospects outweighs huge improvements on the global poverty front. They also often reject similar scope-sensitive arguments for things like animal charities.

Giving What We Can's page on what we can achieve still focuses on global poverty, because developing-world charity is easier to explain persuasively. However, EA leadership tends to privately focus on things like AI risk. Two years ago many attendees at the EA Global conference in the San Francisco Bay Area were surprised that the conference focused so heavily on AI risk, rather than the global poverty interventions they'd expected.

Stage 3: Effective altruism is self-recommending

Shortly before the launch of the EA Funds I was told in informal conversations that they were a response to demand. Giving What We Can pledge-takers and other EA donors had told CEA that they trusted it to GWWC pledge-taker demand. CEA was responding by creating a product for the people who wanted it.

This seemed pretty reasonable to me, and on the whole good. If someone wants to trust you with their money, and you think you can do something good with it, you might as well take it, because they're estimating your skill above theirs. But not everyone agrees, and as the Madoff case demonstrates, "people are begging me to take their money" is not a definitive argument that you are doing anything real.

In practice, the funds are managed by Open Philanthropy Project staff:

> We want to keep this idea as simple as possible to begin with, so we'll have just four funds, with the following managers:
>
> - Global Health and Development - Elie Hassenfeld
> - Animal Welfare – Lewis Bollard
> - Long-run future – Nick Beckstead
> - Movement-building – Nick Beckstead
>
> (Note that the meta-charity fund will be able to fund CEA; and note that Nick Beckstead is a Trustee of CEA. The long-run future fund and the meta-charity fund continue the work that Nick has been doing running the EA Giving Fund.)

It's not a coincidence that all the fund managers work for GiveWell or Open Philanthropy. First, these are the organisations whose charity evaluation we respect the most. The worst-case scenario, where your donation just adds to the Open Philanthropy funding within a particular area, is therefore still a great outcome. Second, they have the best information available about what grants Open Philanthropy are planning to make, so have a good understanding of where the remaining funding gaps are, in case they feel they can use the money in the EA Fund to fill a gap that they feel is important, but isn't currently addressed by Open Philanthropy.

In past years, Giving What We Can recommendations have largely overlapped with GiveWell's top charities.

In the comments on the launch announcement on the EA Forum, several people (including me) pointed out that the Open Philanthropy Project seems to be having trouble giving away even the money it already has, so it seems odd to direct more money to Open Philanthropy Project decisionmakers. CEA's senior marketing manager replied that the Funds were a minimum viable product to test the concept:

> I don't think the long-term goal is that OpenPhil program officers are the only fund managers. Working with them was the best way to get an MVP version in place.

This also seemed okay to me, and I said so at the time.

[NOTE: I've edited the next paragraph to excise some unreliable information. Sorry for the error, and thanks to Rob Wiblin for pointing it out.]

After they were launched, though, I saw phrasings that were not so cautious at all, instead making claims that this was generally a better way to give. As of writing this, if someone on the effectivealtruism.org website clicks on "Donate Effectively" they will be led directly to a page promoting EA Funds. When I looked at Giving What We Can's top charities page in early April, it recommended the EA Funds "as the highest impact option for donors."

This is not a response to demand, it is an attempt to create demand by using CEA's authority, telling people that the funds are *better* than what they're doing already. By contrast, GiveWell's Top Charities page simply says:

> Our top charities are evidence-backed, thoroughly vetted, underfunded organizations.

This carefully avoids any overt claim that they're the highest-impact option available to donors. GiveWell avoids saying that because there's no way they could know it, so saying it wouldn't be truthful.

A marketing email might have just been dashed off quickly, and an exaggerated wording might just have been an oversight. But when I looked at Giving What We Can's top charities page in early April, it recommended the EA Funds "as the highest impact option for donors."

The wording has since been qualified with "for most donors", which is a good change. But the thing I'm worried about isn't just the explicit exaggerated claims – it's the underlying marketing mindset that made them seem like a good idea in the first place. EA seems to have switched from an endorsement of the best things outside itself, to an endorsement of itself. And it's concentrating decisionmaking power in the Open Philanthropy Project.

Effective altruism is overextended, but it doesn't have to be

There is a saying in finance, that was old even back when Keynes said it. If you owe the bank a ***million*** dollars, then ***you*** have a problem. If you owe the bank a ***billion*** dollars, then the ***bank*** has a problem.

In other words, if someone extends you a level of trust they could survive writing off, then they might call in that loan. As a result, they have leverage over you. But if they overextend, putting all their eggs in one basket, and you are that basket, then you have leverage over them; you're too big to fail. Letting you fail would be so disastrous for their interests that you can extract nearly arbitrary concessions from them, including further investment. For this reason, successful institutions often try to diversify their investments, and avoid overextending themselves. Regulators, for the same reason, try to prevent *banks* from becoming "too big to fail."

The Effective Altruism movement is concentrating decisionmaking power and trust as much as possible, in a way that's setting itself up to invest ever increasing amounts of confidence to keep the game going.

The alternative is to keep the scope of each organization narrow, overtly ask for trust for each venture separately, and make it clear what sorts of programs are being funded. For instance, Giving What We Can should go back to its initial focus of global poverty relief.

Like many EA leaders, I happen to believe that anything you can do to steer the far future in a better direction is much, much more consequential for the well-being of sentient creatures than any purely short-run improvement you can create now. So it might seem odd that I think Giving What We Can should stay focused on global poverty. But, I believe that the single most important thing we can do to improve the far future is *hold onto our ability to accurately build shared models*. If we use bait-and-switch tactics, we are actively eroding the most important type of capital we have – coordination capacity.

If you do not think giving 10% of one's income to global poverty charities is the right thing to do, then you can't in full integrity urge others to do it – so you should *stop*.

You might still believe that GWWC ought to exist. You might still believe that it is a positive good to encourage people to give much of their income to help the global poor, if they wouldn't have been doing anything else especially effective with the money. If so, and you happen to find yourself in charge of an organization like Giving What We Can, the thing to do is write a letter to GWWC members telling them that you've changed your mind, and why, and offering to *give away the brand* to whoever seems best able to honestly maintain it.

If someone at the Centre for Effective Altruism fully believes in GWWC's original mission, then that might make the transition easier. If not, then one still has to tell the truth and do what's right.

And what of the EA Funds? The Long-Term Future Fund is run by Open Philanthropy Project Program Officer Nick Beckstead. If you think that it's a good thing to delegate giving decisions to Nick, then I would agree with you. Nick's a great guy! I'm always happy to see him when he shows up at house parties. He's smart, and he actively seeks out arguments against his current point of view. But the right thing to do, if you want to persuade people to delegate their giving decisions to Nick Beckstead, is to *make a principled case for delegating giving decisions to Nick Beckstead*. If the Centre for Effective Altruism did that, then Nick would almost certainly feel more free to allocate funds to the best things he knows about, not just the best things he suspects EA Funds donors would be able to understand and agree with.

If you can't directly persuade people, then maybe you're wrong. If the problem is inferential distance, then you've got some work to do bridging that gap.

There's nothing wrong with setting up a fund to make it easy. It's actually a really good idea. But there is something wrong with the multiple layers of vague indirection involved in the current marketing of the Far Future fund – using global poverty to sell the generic idea of doing the most good, then using CEA's identity as the organization in charge of doing the most good to persuade people to delegate their giving decisions to it, and then sending their money to some dude at the multi-billion-dollar foundation to give away at his personal discretion. The same argument applies to all four Funds.

Likewise, if you think that working directly on AI risk is the most important thing, then you should make arguments directly for working on AI risk. If you can't directly persuade people, then maybe you're wrong. If the problem is inferential distance, it might make sense to imitate the example of someone like Eliezer Yudkowsky, who used indirect methods to bridge the inferential gap by writing extensively on individual human rationality, and did not try to control others' actions in the meantime.

If Holden thinks he should be in charge of some AI safety research, then he should ask Good Ventures for funds to actually start an AI safety research organization. I'd be excited to see what he'd come up with if he had full control of and responsibility for such an organization. But I don't think anyone has a good plan to work directly on AI risk, and I don't have one either, which is why I'm not directly working on it or funding it. My plan for improving the far future is to build *human* coordination capacity.

(If, by contrast, Holden just thinks there needs to be *coordination* between different AI safety organizations, the obvious thing to do would be to work with FLI on that, e.g. by giving them enough money to throw *their* weight around as a funder. They organized the successful Puerto Rico conference, after all.)

Another thing that would be encouraging would be if at least one of the Funds were not administered entirely by an Open Philanthropy Project staffer, and ideally an expert who doesn't benefit from the halo of "being an EA." For instance, Chris Blattman is a development economist with experience designing programs that don't just use but generate evidence on what works. When people were arguing about whether sweatshops are good or bad for the global poor, he actually went and looked by performing a randomized controlled trial. He's leading two new initiatives with J-PAL and IPA, and expects that directors designing studies will also have to spend time fundraising. Having funding lined up seems like the sort of thing that would let them spend more time actually running programs. And more generally, he seems likely to know about funding opportunities the Open Philanthropy Project doesn't, simply because he's embedded in a slightly different part of the global health and development network.

Narrower projects that rely less on the EA brand and more on *what they're actually doing*, and more cooperation on equal terms with outsiders who seem to be doing something good already, would do a lot to help EA grow beyond putting stickers on its own behavior chart. I'd like to see EA grow up. I'd be excited to see what it might do.

Summary

1. Good programs don't need to distort the story people tell about them, while bad programs do.
2. Moral confidence games – treating past promises and trust as a track record to justify more trust – are an example of the kind of distortion mentioned in (1), that benefits bad programs more than good ones.
3. The Open Philanthropy Project's Open AI grant represents a shift from evaluating other programs' effectiveness, to assuming its own effectiveness.
4. EA Funds represents a shift from EA evaluating programs' effectiveness, to assuming EA's effectiveness.
5. A shift from evaluating other programs' effectiveness, to assuming one's own effectiveness, is an example of the kind of "moral confidence game" mentioned in (2).
6. EA ought to focus on scope-limited projects, so that it can directly make the case for those particular projects instead of relying on EA identity as a reason to support an EA organization.
7. EA organizations ought to entrust more responsibility to outsiders who seem to be doing good things but don't overtly identify as EA, instead of trying to keep it all in the family.

(Cross-posted at my personal blog and the EA Forum.

Disclosure: I know many people involved at many of the organizations discussed, and I used to work for GiveWell. I have no current institutional affiliation to any of them. Everyone mentioned has always been nice to me and I have no personal complaints.)

Project Hufflepuff: Planting the Flag

> *"Clever kids in Ravenclaw, evil kids in Slytherin, wannabe heroes in Gryffindor, and everyone who does the actual work in Hufflepuff."*
>
> *- Harry Potter and the Methods of Rationality, Chapter 9*
>
> *"It is a common misconception that the best rationalists are Sorted into Ravenclaw, leaving none for other Houses. This is not so; being Sorted into Ravenclaw indicates that your strongest virtue is curiosity, wondering and desiring to know the true answer. And this is not the only virtue a rationalist needs. Sometimes you have to work hard on a problem, and stick to it for a while. Sometimes you need a clever plan for finding out. And sometimes what you need more than anything else to see an answer, is the courage to face it...*
>
> *- Harry Potter and the Methods of Rationality, Chapter 45*

I'm a Ravenclaw and Slytherin by nature. I like being clever. I like pursuing ambitious goals. But over the past few years, I've been cultivating the skills and attitudes of Hufflepuff, by choice.

I think those skills are woefully under-appreciated in the Rationality Community. The problem cuts across many dimensions:

- Many people in rationality communities feel lonely (even the geographically tight Berkeley cluster). People want more (and deeper) connections than they currently have.
- There are lots of small pain points in the community (in person and online) that could be addressed fairly easily, but which people don't dedicate the time to fix.
- People are rewarded for starting individual projects more than helping to make existing ones succeed, which results in projects typically depending on a small number of people working unsustainably. (i.e. a single person running a meetup who feels like if they left, the meetup would crumble apart)
- Some newcomers often find the culture impenetrable and unwelcoming.
- Not enough "real-time operational competence" - the ability to notice problems in the physical world a and solve them.
- Even at events like EA Global where enormous effort is put into operations and logistics, we scramble to pull things together at the last minute in a way that is very draining.
- Many people communicate in a way that feels disdainful and dismissive (to many people), which makes both social cohesion as well as intellectual understanding harder.
- We have a strong culture of "make sure your own needs are met", that specifically pushes back against broader societal norms that pressure people to conform. This is a good, but I think we've pushed too far in the opposite direction. People often make choices that are valuable to them in the immediate term, but which have negative externalities on the people around them.

In a nutshell, the emotional vibe of the community is preventing people from feeling happy and and connected, and a swath of skillsets that are essential for group intelligence and ambition to flourish are undersupplied.

If any one of these things were a problem, we might troubleshoot it in isolated way. But collectively they seem to add up to a cultural problem, that I can't think of any way to express other than "Hufflepuff skills are insufficiently understood and respected."

There are two things I mean by "insufficiently respected":

- Ravenclaw and Slytherin skills come more naturally to many people in the community, and it doesn't even occur to people that emotional and operational skills are something they should cultivate. It feels like a separate magisteria that specialists should do. They're also quick to look at social niceties and traditions that seem silly, make a cursory attempt to understand them, and then do away with them without fully understanding their purpose.
- People who might join the community who value emotional and operational skills more highly, feel that the community is not for them, or that they have to work harder to be appreciated.

And while this is difficult to explain, it feels to me that there is a central way of being, that encompasses emotional/operational intelligence and deeply integrates it with rationality, that we are missing as a community.

This is the first in a series of posts, attempting to plant a flag down and say "Let's work together to try and resolve these problems, and if possible, find that central way-of-being."

I'm decidedly not saying "this is the New Way that rationality Should Be". The flag is not planted at the summit of a mountain we're definitively heading towards. It's planted on a beach where we're building ships, preparing to embark on some social experiments. We may not all be traveling on the same boat, or in the exact same direction. But the flag is gesturing in a direction that can only be reached by multiple people working together.

A First Step: The Hufflepuff Unconference, and Parallel Projects

I'll be visiting Berkeley during April, and while I'm there, I'd like to kickstart things with a Hufflepuff Unconference. We'll be sharing ideas, talking about potential concerns, and brainstorming next actions. (I'd like to avoid settling on a long term trajectory for the project - I think that'd be premature. But I'd also like to start building some momentum towards some kind of action)

My hope is to have both attendees who are positively inclined towards the concept of "A Hufflepuff Way", and people for whom it feels a bit alien. For this to succeed as a long-term cultural project, it needs to have buy-in from many corners of the rationality community. If people have nagging concerns that feel hard to articulate, I'd like to try to tease them out, and address them directly rather than ignoring them.

At the same time, I don't want to get bogged down in endless debates, or focus so much on criticism that we can't actually move forward. I don't expect total-consensus, so my goal for the unconference is to get multiple projects and social experiments running in parallel.

Some of those projects might be high-barrier-to-entry, for people who want to hold themselves to a particular standard. Others might be explicitly open to all, with radical inclusiveness part of their approach. Others might be weird experiments nobody had imagined yet.

In a few months, there'll be a followup event to check in on how those projects are going, evaluate, and see what more things we can try or further refine.

[Edit: The Unconference has been completed. Notes from the conference are here]

Thanks to Duncan Sabien, Lauren Horne, Ben Hoffman and Davis Kingsley for comments