

[The authoritative copy of this document can now be found on the EA Forum.](#)

Toward Impact Markets

Summary

We want to create a market to trade “impact certificates” – a variation on [social impact bonds](#).

The backbone of the market are retroactive funders who are transparent about the sorts of impact certificates they want to see and buy them when they are offered. They usually only start to be interested in impact certificates when they describe actions that are in a mature state with a low remaining risk of failure. Some organizations will be in a position to complete such actions without help, but others will need to fundraise or partner with others first. An impact market is where this fundraising and partnering can take place. Anyone can use it to speculate on the retroactive funding.

Impact markets promise to partially solve a range of pressing problems that altruists face. (See sections Retroactive Funding and Benefits.) Nonetheless they bear a number of benign risks that can cause them to fail and more dangerous risks that can cause them to be harmful. (See section Risks.) We think it will be possible to address the risks in the second category, but we would like to be more certain of that. (See section Solutions.) Small-scale, sandboxed experiments among highly informed participants may help us test impact markets while limiting their risks, which is what some parties are now attempting. (See section Current Work.)

We hope that this article will give readers a chance to critique our plans and point out further risks or weaknesses in our defenses against known risks. If you would like to help, [please join our Discord](#), comment, or contact us some other way.

Impact Markets

Our current best guess is that impact markets will take the shape of four incremental additions to existing financial markets.

1. **Retroactive funding.** First there needs to be a norm of and ideally several institutions for retroactive funding. They have a lot of responsibility for the health of the market and need to monitor it closely. We don't think that it's feasible for them to operationalize up front exactly what the impact looks like that they want to buy (say, by pointing to particular metrics) but rather want to rely on indirect normativity: Market participants can read about them and talk to them and try to understand their values. Ideally there are many retro funders who are all highly qualified and responsible and who are getting more numerous over time and whose capital increases. Some of these retro funders may have been started with the explicit goal to incentivize more retro funders by retro-funding existing retro funders. (We'll use the terms “retro funder” and “to retro-fund” as convenient short forms.)
2. **Contracts.** Anyone who founds, funds, or otherwise supports a project – a supporter – and wants to participate in its success needs to negotiate their inclusion in a contract (an *impact certificate*) that

documents what share of the impact of the project they own. Retroactive funders will later require this proof if they are interested in buying the impact from the supporter. (Whether it's possible to produce such proof retroactively depends on the specifics of the case.)

3. **Auction platform.** Seed investors are probably hesitant to invest into impact if they have to fear that it'll be hard or take a long time for them to find a buyer for it. An auction platform can streamline the interaction between the buyers and the sellers of impact.
4. **Impact stock and derivatives.** Finally, at some later date, charities may decide to create their own stock. Alternatively, some third parties may create derivatives that track the market capitalization of all impact certificates of one charity. These will make some forms of investment easier, and charities could use them for mission hedging.

Notes

1. This whole document culminates in a particular definition of *Attributed Impact*. All mentions of "Attributed Impact" and just "impact" refer, by default, to Attributed Impact according to that definition.
2. What we call "charity" is a charity in spirit but may be an Impact DAO or an entity incorporated as for-profit if that is necessary. Examples: Against Malaria Foundation, Wave, Protocol Labs.
3. What we call a "project" is a precisely defined undertaking of, for example, a charity. Examples: one distribution of long-lasting insecticide-treated bednets, an epic in a software development process, a conference.
4. We want impact markets to be interesting for (1) altruistic consequentialists and (2) profit-oriented capitalists. (Most people are a bit of both.) If it ends up being useful for collectors too, all the better, but we want impact markets to work regardless of whether collectors get interested in them.

Retroactive Funding

Retroactive funders benefit from their actions in various ways.

Reduced Uncertainty

Retroactive funders only fund projects that have achieved a certain degree of success.

Projects usually go through two phases, a phase where finding out whether it'll be successful is a question that management experts and priorities research experts need to collaborate on and a phase where the management experts are not needed anymore.

Let's say that the project is a proposed conference a year in the future.

In the first phase the project can fail because the team behind it splits up, because the team procrastinates until all venues are booked, or because a pandemic strikes. It can also fail because the attendees come away excited about Homeopaths Without Borders or because one attendee took out vacation days to attend and their replacement accidentally triggered a Dead Hand mechanism.

The second phase starts when the conference is over, the venue is cleaned, and the feedback from the attendees is in. In this second phase all the uncertainty concerning the team, their procrastination, and the pandemic is gone. What remains is just the second type of uncertainty. (Or some part of it since the clumsy

replacement has probably returned to their accustomed job away from the Dead Hand mechanism.) The uncertainty at this stage is strictly lower than the uncertainty during the first phase.

Let's suppose that a retroactive funder has the requirement on a conference that it must've happened, that not too many attendees died from Covid, and that the feedback was at least 90% positive. Let's further suppose that they would have to spend one day per proposed future conference to assess whether the team behind it will reach that bar, and that then they'll be right in 1 in 5 cases.

Not having to do all of this means saving 5 times 1 day of work per successful conference for the evaluation and keeping the grant money for longer. So even if they buy the impact of the 1 in 5 successful conferences at 5× the price, they still make a profit, altruistically speaking.

Reduced Evaluation Time

Let's suppose that the funder supports a lot of conferences that follow roughly the same pattern, so that can be said to implement the same intervention, and that the funder has invested enough time up front to now be quite confident in the value of the intervention.

If the 5 days for the team evaluation are half the time that they would, counterfactually, spend per project (counting the up-front investment into finding the intervention), then they can buy the successful conferences at $(2 \times 5) \times$ the price of the counterfactual grant and still make a profit. (The factor 5 is from the subsection above.)

Longer Market Exposure

Since they invest later, they can keep their money invested in yield-generating assets for longer. In most cases this will have a small influence, say, $1.1 \times$ per year.

Earning a Double Bottom Line

But there are even more benefits once impact markets are better established. Some retroactive funders may be quite conservative with their investments, so that they don't think that their retroactively purchased impact certificates are (in impact terms) hundreds of times more valuable than the average impact certificate that gets purchased by other retroactive funders. These retroactive funders may be interested in doubling as a speculator in other parts of the impact market.

Usually they'd have to park their money for a few years, decades, or centuries in classic stocks, bonds, or ETFs, which generate comparatively little impact just to maintain their liquidity and thus option value. With impact markets that becomes unnecessary as they can put money in charities (essentially becoming investors too) and then take it out again when they want to use it for their retroactive funding.

As retroactive funders themselves they may also have the expertise to predict what other retroactive funders will want to buy, so that they can be among the most successful investors on the market. But we'd still think that the social bottom line will do the heavy lifting in that portfolio, so that it's probably not so attractive for retroactive funders who think that an actual extra dollar for their retroactive funding is worth a lot more than a dollar to one of another retroactive funder's favorite charities. For example, it might be that they estimate that their monetary profits will shrink by 1 percentage point but that the social bottom

line will be similar to what the Against Malaria Foundation would have achieved with ten times that money. A GiveWell-type funder would be excited about that investment opportunity while a donor who typically supports organizations like MIRI would want to avoid it.

Finally, retroactive funders may burn their shares in impact certificates if speculators start avoiding them because they fear that the supply is too great. But if that is not a worry, other future retroactive funders may also just buy shares from earlier retroactive funders in a never-ending chain of retroactive funding. Such a chain will also signal that the retroactive funding is likely to only ever increase, which would boost long-term investment into impact certificates.

The gains from this parking are hard to pin down. The worst-case is close to $1\times$ since it's optional and funders won't do it if it's not useful for them. In the other direction it's probably capped at around $20\times$ since retroactive funding would not be attractive in worlds in which seed funding is (still) that effective even on the margin.

Spotting Blindspots for Funding Opportunities

If the space is big with hundreds of promising charities with dozens of certificates each at any given time, then funders may be overwhelmed and miss great retro-funding opportunities if they are outside their social circles, in another country and language, or even just very unusual. A market made up of speculators from many different industries and countries may be better at recognizing such niche opportunities than any one funder. They will also be incentivized to make sure that the funder knows about them.

Conversely, retro funders need to be wary of the temptation to ignore unpopular impact certificates. Rewarding the sorts of speculators who are smart enough to notice great funding opportunities that no one else notices is exactly the sort of mechanism that makes impact markets valuable for prioritization.

Contracts (Impact Certificates)

Impact certificates, in our model, are contracts that regulate how to delimit and distribute some unit of change to the supply of a public good. (See below for the Attributed Impact definition that refines this “change to the supply of a public good.”) Owen Cotton-Barratt and we plan to publish a set of rules that will provide more clarity here. (You can find a sneak peek in our [Funding the Commons talk](#).) They aim to smooth out some initial friction that a marketplace for impact certificates may face. We'll summarize them briefly in the following.

The guiding metaphor. A winery is typically credited with producing wine (which corresponds to impact). This assumes that employees, letters of the land, states, ancestors, meteorologists, insects, et al. forfeit their claim to the wine one way or another, either by custom or through a contract. Wine can flow freely or evaporate, but it can also be bottled and sold. (The full bottle corresponds to an impact certificate.) The bottling, in turn, can be undone, which is usually followed by the consumption of the wine. That consumption, we assume, is final.

This metaphor should give a rough overview of the spirit of the rules.

The rules propose that impact certificates be:

1. as clear and concrete as possible, and
2. issued only by all those actors collectively who own classic legal rights in that which created the impact.

They also recommend that the market system make them:

3. consumable in the sense that one can permanently reverse the issue, and
4. incremental in the sense that they don't require changes to existing financial markets.

Generally, we think that markets may well home in on these norms over time by themselves, but we're currently in a good position to prevent all the frictions that would come with that. We think that in particular the first two rules are crucial. We're more on the fence about whether rules 3 and 4 are really needed.

The rest is commentary:

The first two rules jointly limit the kinds of outputs that can be bottled up in impact certificates. The limits are not hard, but greater ambiguity will make it harder to find buyers for an impact certificate: Buyers of overly vague impact certificates may make losses if others later sell impact certificates that seem to substantially overlap. No one will quite know who owns the overlapping bits, and bid prices will drop. These buyers will later take good care not to buy overly vague certificates.

Ambiguity can have two sources: confusion over what is being sold and confusion over who (else) might make claims to that which is being sold. We therefore advise issuers of impact certificates to word them carefully and to make contracts in writing with whoever might make a claim on the impact. Note that such contracts needn't specify a concrete fractional allocation but can also specify an algorithm (such as [SourceCred](#)) that is to be used to determine the allocation at the time of the issue.

Eventually, we imagine, aggregator and auditor firms will emerge that are independent of any issuers. Aggregators will compile all information on all impact certificates that have ever been issued on any market so that duplicates can be exposed. Issuers can then get audits to prove to buyers that they are honest and are not trying to double-spend their impact. (For example, someone might first sell their impact from "A fundraiser for Rethink Priorities," and later sell their impact from "A fundraiser at a vertical farming expo" when really those are about one and the same fundraiser.)

Examples. A good definition may sound broadly like, "Within the 264th year of the era of Aelius Antoninus, I, Hypatia, will ghost-write and publish a paper on behalf of Ada Lovelace that proves (once she is born) that Vingean reflection is possible. I'm selling half of the impact. The other half will belong to Ms. Lovelace. There are no funders outside the market, and there are no conflicts of interest to disclose. In fact, our interests will be similar."

(Or with reduced specificity: "Within the year N, I, Person A, will publish a paper that proves X, coauthored with Person B. I'm selling half of the impact. The other half belongs to Person B. There are no funders outside the market, and there are no conflicts of interest to disclose.")

A paper has a clear owner, there are no funders who don't participate in the market (including Hypatia herself), it is clear what she'll do, how long it will take, and that afterwards the paper will be public, and it is clear that the uncertainty over the deeper merits of the work will take many centuries to be resolved. (Ada Lovelace lived about 1,400 years after Hypatia.) All this certainty and uncertainty can be priced in by the market.

A middling definition may sound like, “I, Eve, will distribute 1,000 copies of the attached Vegan Outreach leaflet at Barbican Station in London between April 1 and August 1, 2022. Vegan Outreach has waived all claims on the impact. I will sell 90% of the impact and retain 10% to appease anyone who has switched to lower-suffering behaviors in response to the leaflet but is not content to yield all of their impact to me.”

This definition is highly concrete but there is no clear concept of ownership or rights or responsibility to ground it in. As a result, some of Eve’s newly made vegans who have started to go leafleting themselves may want to sell their impact from their own leafleting. The valuation of Eve’s certificate would have to be enormous to appease them all with just 10% of it. Speculators should price in that there’s a lot of fuzziness here about how much of the impact is owned by who.

A really bad definition is something like, “I, Hancock, will use my superpowers to do something for animal rights.” It is vague in almost all the ways it can be vague. There is a broad idea of a strategy in there, but speculators will have no idea whether they should price it like a corporate campaign or like leafleting. Worse, the definition also leaves the door open to very harmful activities, so that, regardless of the actual outcomes, the Attributed Impact is likely negative. A less extreme version of this is a hypothetical impact certificate for a whole organization that is still active or the whole life of a live person. Any prediction as to how these may change their strategies over the coming decades will have high variance so that the Attributed Impact will almost inevitably be low.

The third rule, the consumption mechanism, allows certificate owners to influence prices by verifiably signaling that they will never sell the certificate. Possible tax-exemptions could be tied to consumption, which may be a minor concern for most investors, but some may invest through a legal entity that can only make tax-exempt grants. The consumption mechanism has also been called “dedication” and “burn.” We’re not convinced that this mechanism is necessary at first, but time will tell. One indicator that it is necessary will be if investors don’t expect impact certificates to be sufficiently deflationary.

Tax exemption. Note that some organizations in the crypto space seem to function like (our phrasing) “optional donor advised funds” in that they have tax exemptions in various countries and can write donation receipts but will only do so if the donor chooses to donate the money, which they don’t have to do. A retro funder could interact with the market through such an entity. If they choose to resell their impact certificates, they will not receive a donation receipt, but when they choose to burn their certificate, they get it. The donation receipt will of course then be from that intermediary rather than from the charity that actually received the money.

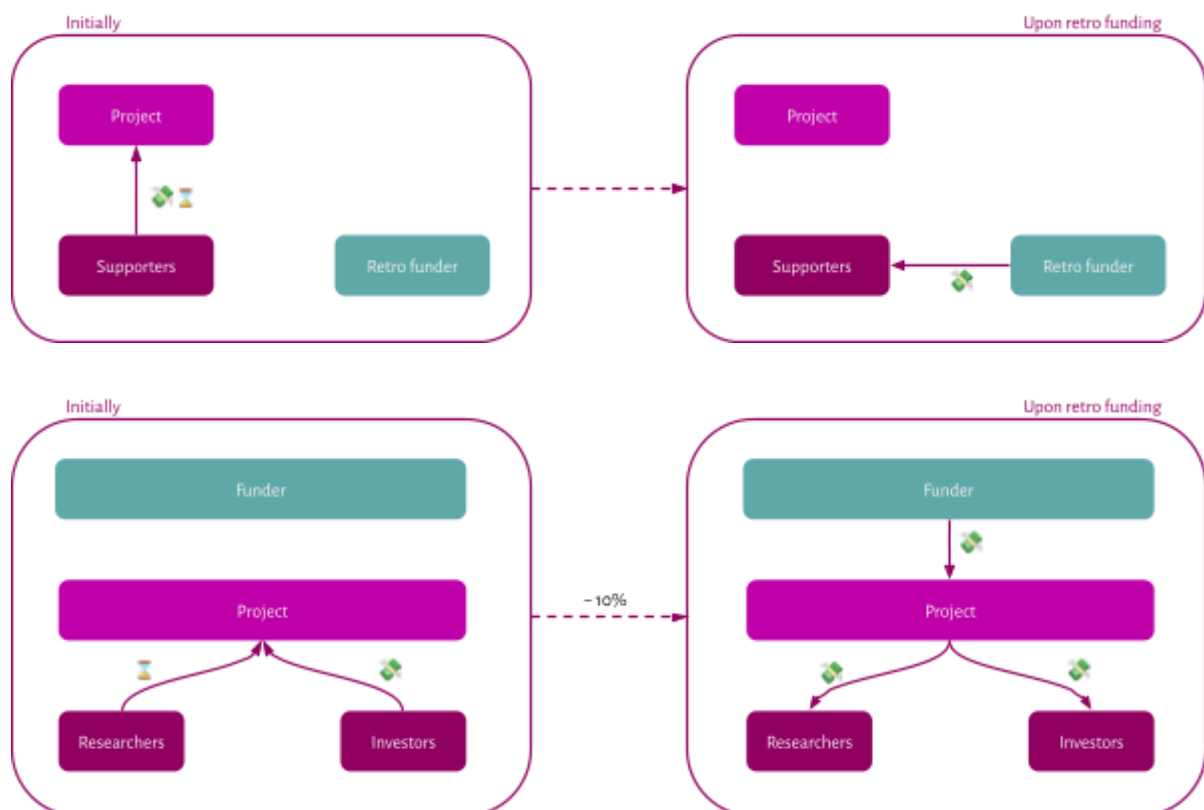
The fourth rule aims to allay worries that impact markets, if successful, will cause revolutionary changes to existing allocation mechanisms. Avoiding such changes avoids opposition, which should make it much easier to institute impact markets. Conversely, even fervent supporters of impact markets may find proposals unrealistic that require changes to existing, established markets.

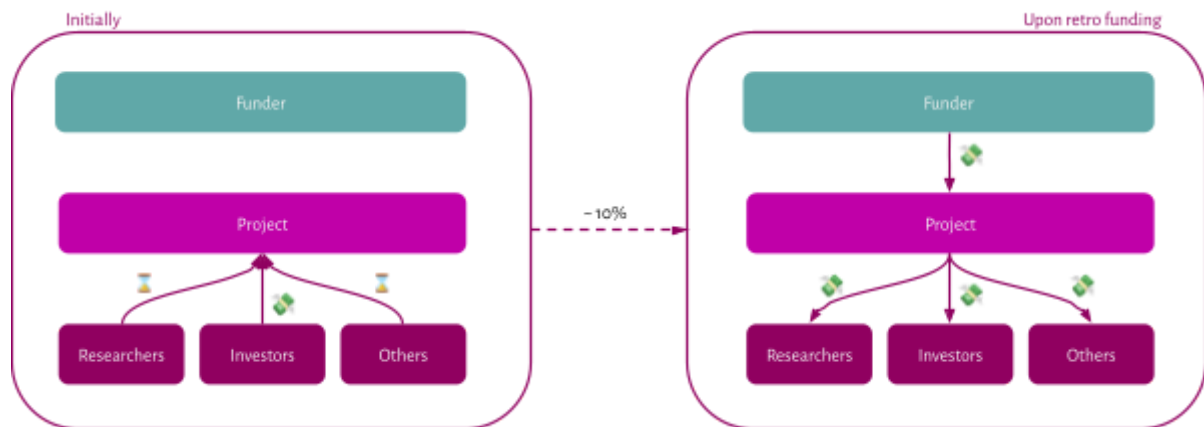
Impact Market Participants

The key idea behind impact markets is that we want to reward supporters (such as founders, investors, and advisors to charities) who proved a particular aptitude for predicting what interventions will later be seen as impactful by sophisticated altruists.

There are three types of fundamentals on an impact market:

1. **Projects:** These are verifiable, clearly ownable, and clearly delimited sets of actions that are detailed in an impact certificate. They benefit from:
 - a. Seed funding,
 - b. More options for aligning incentives.
2. **Supporters:** These include founders, advisors, seed funders, et al. of a charity. Everyone who negotiates a share in the impact certificate. We also use the terms *investor* and *speculator* for the seed funders. They benefit from:
 - a. The seed funding of the project (founders and employees),
 - b. Being able to make an exit (founders and other shareholders),
 - c. Passive income (investors).
3. **Retroactive funders:** The backbone of the market. The funders who buy impact certificates once the projects have matured to a point where the retroactive funders are confident enough in their evaluation of the projects' impact. They reward all supporters of a successful project generously, because they are thankful for not having to bear the risks of all the projects that fail even by their own lights. They benefit from:
 - a. Not carrying the risk for failing projects (5–10x monetary savings),
 - b. Not having to do charity evaluation (team, market, etc.) in addition to intervention evaluation (2x time savings),
 - c. Investing later in time (1.5x monetary savings),
 - d. Parking money (in the role of a speculator) where it generates a double bottom line (1–20x savings?),
 - e. Being able to draw on the market to notice blindspots in the in-house prioritization.





Let's suppose the Against Malaria Foundation (AMF) wants to do a distribution of long-lasting insecticide-treated bednets with its old distribution partner Concern Universal (CU). Let's further suppose that it wants to use impact certificates for this and not its usual funding channels.

(This is an example only. We didn't talk to AMF or CU about any of this and don't actually want any charities that don't have expertise in innovative financial products to become involved so early in the process!)

1. AMF ascertains that some retroactive funders are still interested in more net distributions.
2. The AMF staff decide how they want to split the Attributed Impact from the distribution among themselves, the legal entity AMF, all their partners, contributors, and advisors, and their funders. These decisions can be adjusted in later negotiations. (we'll treat AMF, the legal entity, and all its staff as just "AMF" in the following to save space.)
3. AMF decides on the details of the auction process, that is, what fraction of the profits from each sale should go to AMF and what the minimum percentage raise is between bids.
4. AMF writes an impact certificate.
5. AMF contacts CU, negotiates their split with them, and sends them their fraction of the impact certificate. (we'll assume here that the distribution partner has its own source of seed funding just to make its case interestingly different from AMF's.)
6. AMF contacts a venture capital firm. The VC agrees to fund the distribution. They negotiate the split, and AMF sends them their fraction of the impact certificate.
7. A year later the distribution gets completed successfully.
8. A retro funder notices the successful distribution, is ecstatic, and buys fractions of the certificate from all holders. AMF has no work with this: Any holder just has to give their certificate (or a fraction thereof) to the buyer.

What we seek to demonstrate here is that impact certificates (the way we construe them) are not magic. They don't need to be imbued with any deep meaning beyond that of any other contract.

Auction Platform

We want to create an auction platform to streamline the whole process. ([You can keep up-to-date on the progress here.](#))

Our ideas for MVPs differ in two main ways: (1) The MVP can either emulate the classic market mechanisms where multiple shareholders co-own a project, or it can use the Harberger tax auction where the original

issuer is paid from the profit of each sale, sales are forced, and fractionalizing ownership among multiple shareholders is optional; and (2) it can be aimed at a crypto audience and run on a blockchain, or it can be aimed at a non-crypto audience, not use a blockchain, and avoid monetary transfers. (Note that we use auction and market synonymously here.)

Either way, the critical mechanism is that there's someone (a retro funder) who rewards earlier supporters (founders, funders, advisors, et al.) if they made what has turned out to be good calls.

The classic, non-Harberger auction (or market) has the advantages that:

1. **Alignment.** You can align incentives with cofounders, employees, advisors, partner organizations, et al. by giving them shares in the impact.
2. **Seed funding.** You can get investments by selling parts of the impact to early investors.
3. **Exit.** You can participate in the success of your project by keeping some fraction of the impact until the project has come to fruition (by your assessment) and only then exit. As a founder you will be unusually convinced of it (because if more people were, it would typically already exist), and by extension, you will be more optimistic about it than your investors. Hence you'll want to retain as much of your impact as you can afford until the day has come when you have the proof that will convince the investors too.
4. **Scaling down.** You can start big projects by getting several investors who each buy smaller shares of the impact. Conversely, even smaller investors can get exposure to big projects by buying small shares in them.

The Harberger auction shares two of them even without shared ownership:

1. **Alignment.** You can align incentives with cofounders, employees, advisors, partner organizations, et al. by making separate contracts with them that they'll receive part of the profit share. That would be similar to an ordinary participation.
2. **Seed funding.** You can get seed funding by selling all of your impact to early investors. You can set the minimum bid so that the funding is enough to bootstrap the project, but you can't have any further fundraising rounds.
3. **Simplicity.** You don't need to countertrade your investors once you have the seed funding. You'll simply profit from each of their sales automatically.

With shared ownership, the Harberger auction shares all the benefits of the classic system, but in practice there will be trade-offs:

1. A larger issuer share (or tax) will discourage speculation because it's subtracted from the profits of the speculator. The market will be less liquid as a result. But issuers can set an arbitrarily small share to manage that risk. [Crypto exchanges](#) seem to work well with taker fees around and slightly below 0.05%.
2. But a small share also means small profits. Prices can't decrease in the Harberger auction, so the profits will be proportional to the price. (If they could decrease, they'd be proportional to the volume, and charities had an incentive to create volatility.) In the classic system, you and your collaborators may collectively manage to retain 10–30% of your impact until you exit, which you have to time such that you don't sell too early. In the Harberger system you don't have to worry about the timing, but you'll all collectively only “retain” (through the tax) 0.05% unless you want to

sacrifice market liquidity. Even if you choose to set the tax at 1%, that's still 10x the cost, plus the reduced liquidity, in exchange for simplicity.

All in all, a Harberger auction with fractional ownership seems like it maximizes option value because issuers can choose to turn it into a classic auction by setting their tax to 0.

The current state is as follows:

Blockchain (web3). Our first testing ground is a [smart contract on an Ethereum test net](#) that implements an auction as described above. It does not currently support fractional impact certificates, but we think that that can be added in the form of a separate third-party smart contract. In general, a blockchain-based or web3 solution would have a number of advantages:

1. It makes it easy to interface with the existing web3 ecosystem around public goods funding.
2. The existing ecosystem is itself a great testing ground for innovative solutions because everyone is unusually knowledgeable of innovative financial tools, so we're at less of a risk of making mistakes that would cause other people to lose money.
3. It allows for great scale as transactions are fast and frictionless, and, depending on the blockchain, also cheap.
4. It makes it easier to market the solution because users don't need to trust us if they can read the code and there are audits.

(You can watch a demo in [our talk at Funding the Commons II.](#))

Spreadsheet (web2). Another MVP that we have in mind is a spreadsheet (via Google Sheets) that tracks transactions between issuers, speculators, and retroactive funders (any pairwise permutations). One sheet simply records the transactions; another summarizes the transactions such that they show the latest configuration of ownership; and a third sheet displays some statistics, such as the price history, market capitalization, volume, etc. We haven't created this spreadsheet yet, but we expect that it'll be hard to scale for various reasons related to error-reporting and because spreadsheets get slow when there's a lot of data in them.

Crucially, the spreadsheet would not execute the monetary transactions themselves, it would just record them. A transaction is then said to be pending when one side has entered it but the other hasn't confirmed it. It is only considered for the statistics once both sides have confirmed it. Google Sheets allows owners to protect ranges and individual cells such that only one person can edit them, so that aspect should be fairly failsafe.

A variation on the same theme is a spreadsheet that only tracks what projects accept seed funding and what projects have been completed. All seed-funding transactions could then happen over a platform like Kickstarter or Indigogo. Once the project is completed, a retro funder can consider it and transfer their money either directly to all contributors or to the project founder so that they can forward it to all contributors.

We could test the spreadsheet solution by allowing people to use it to trade on retro funding for something very safe and easily verifiable such as EA Forum articles. One could also call this a "proportional prize pool for prescient philanthropists" because they are basically prizes for people who've made good calls – as founders, funders, supporters, or similar – except that not only the top 3 or top 10 get prizes but almost

everyone who clears some minimal bar of the retro funder and has invested enough that the prize is worth the transfer overhead.

A web2 solution like that would have a few advantages too:

1. It is quick to iterate on as it doesn't require audits. It won't handle financial transactions, and even if it did from the users' perspective, they would really be handled by a payment gateway like Datatrans.
2. It would not handle financial transactions, so many legal risks from those shift away from us to the users of the system, who we'll need to warn to do their research.
3. It'll be more frictionless for the sorts of users who haven't used web3 solutions before.

Personally, we're also a bit annoyed by people talking so much about valuing impact certificates in and of themselves like that's important, so we like solutions that don't obviously contain any one component that is an impact certificate. (we see an impact certificate like any other contract that we value only for the monetary or altruistic bottom line that it nets "us" – scare quotes because only the monetary bottom line is agent-relative.)

We are fairly convinced that the blockchain-based solution is going to be the culmination of our efforts one day, but we're ambivalent over which MVP will allow us to test the market more quickly and productively.

Impact Stock and Derivatives

At later stages, we imagine that investors will find it overly complicated to use individual auctions to bid on countless projects. Some charities may also prefer to use other kinds of auctions at least for some of their projects. Plus, the trust that a project will be impactful is probably going to be tied to the team behind it, the charity, so that people, especially those who are lay people when it comes to the interventions of the charity, will prefer to invest into charities instead of individual projects.

We imagine that this can be achieved with derivatives (such as perpetual futures) that track the market capitalization of all impact certificates that a charity has issued. Alternatively, and if that is legally permissible for the charity, it could issue its own stock, which it could also use to hedge against volatility in the flow of donations that it gets through other channels.

Benefits

(We briefly covered this section in [our talk at Funding the Commons II](#).)

We think that impact markets are most suitable for hits-based funding for individuals or young charity startups without impressive track records. [This article makes this argument in greater detail](#). In short, reference classes of projects where success probabilities over 80–90% are common make it hard for investors to think that they have enough private information about a particular project that they will accept the low success-conditioned rewards that retro funders will pay in such cases.

The following table gives an overview – i.e. rough tendencies based on broad simplifications – of who will find which benefit to be interesting:

	Funders	Charities
Greater maximum scale	✓	–
Faster scaling	–	✓
More funding opportunities	✓	–
Better access to funding	–	✓
Greater hiring pools	✓	✓
Incentives for excellence	–	✓
More priorities research	✓	–

Benefits for

Greater Maximum Scale

Minor Benefits

Funders can probably always benefit from the greater liquidity so long as there is any impact market in any space that they value enough that the profits plus the social bottom line are more valuable to them than the profits they would've gained from their conventional financial market investments. That is probably always true (given the existence of liquid impact markets) if a funder is rather conservative. Purely hits-based funders may prefer their retro funded projects over average projects by such an enormous margin that the social bottom line of a speculative investment will never sway the balance for them.

Note that I've bracketed "priorities research & diversity" in the case of the conservative funder in a big space. A conservative funder is, to me, at first approximation, a funder who optimizes for expected value (just like the hits-based one) under the constraint that the probability of success needs to be much greater than 10%. In that case they would benefit from "priorities research & diversity." But *de facto* my impression of conservative funders is that they optimize for the probability of success with very little regard to expected value at all. There just needs to be some sort of minimal-bar-meeting success, like a debriefing that shows that something happened. In that case they can probably find enough funding opportunities that meet their bar in a big space without any help from investors.

Aligning Incentives

Charities typically like to hire very closely value-aligned people. That means that they can pay lower salaries because everyone cares about the mission. But that doesn't seem like the optimal state. There are people who care about the mission but also have a family to feed or a debt to repay. There are also extremely capable people who care a bit less about the mission. The charity will lose out on them.

If we now assume that the employee that the charity wants to hire has a bit of runway and can run about as much risk as an early startup employee. If they think that a retroactive funder will like the charity enough

to buy its impact certificates, that employee may agree to a deal where the regular salary plus the expected value of 1% of all impact certificates adds up to more than a regular salary. If the employee cared a bit about the charity's mission before, now they care a lot.

Incentivizing Excellence

Many funders today aim to fill or to partially fill the funding gaps of any project that clears some bar in terms of expected cost-effectiveness. So highly capable startup founders have a tradeoff to make whether they want to start a for-profit business and donate billions or whether they want to start a charity and get at most exactly as much as they need. They wouldn't want more unless they have enough spare time to become better grantmakers than those who make grants to them. But if charities can fundraise from for-profit venture capitalists, the gulf between charities and businesses shrinks. This means entrepreneurs have a less difficult tradeoff to consider when choosing between launching a charity or a for-profit.

The quick, quantitative feedback could also by itself be motivational for founders and employees.

Creating Cheaper Liquidity for Charities and Funders

Large funders currently park their money in for-profit businesses until the growth rate from the growing wisdom of the funder (plus the low average growth rate of the businesses) drops below the growth rate of the public goods. That can take a long time and is therefore very wasteful, but it's a necessary evil because investments into public goods are currently illiquid, so if you get them wrong, you can't withdraw the money again. Impact markets would change that, and large funders could park their money in ETFs that comprise many big, well-established and somewhat impactful charities until they find better investment vehicles.

Consequently many charities will receive funding more quickly and liberally and greater amounts of it because it is not as costly in terms of option value for the funder.

Scaling Priorities Research

It's very hard to assess what will have a positive impact. It's somewhat less hard to assess what has had a positive impact. If funders concentrate on getting the second as right as they can, they can outsource the first to a market where investors who would not otherwise spend their time on altruistic concerns will grapple with it.

Aggregating a Diversity of Perspectives

Markets aggregate information from many investors. These investors might all come from the same social bubbles and so share similar knowledge and knowledge gaps, but that problem will be less pronounced than for individuals or fixed grantmaking teams. So a market will be able to improve on the breadth of information that we draw on when we prioritize between interventions.

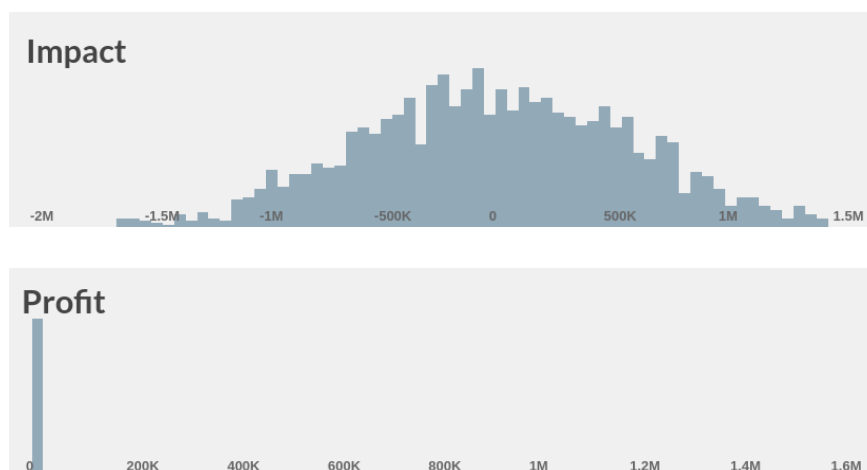
Risks

(We briefly covered this section in [our talk at Funding the Commons II.](#))

Impact and Profit Distribution Mismatch

Investors should have strong reasons to expect that the prices of certificates will, in the limit, be proportional to the value that a [Pareto-optimal compromise axiology](#) would assign to them – that is the moral standard that is reached only when no gains from moral trade are left on the table.

But we think that is unlikely to happen by default. There is a mismatch between the probability distribution of investor profits and that of impact. Impact can go vastly negative while investor profits are capped at only losing the investment. We therefore risk that our market exacerbates negative externalities.



Standard distribution mismatch. Standard investment vehicles work the way that if you invest into a project and it fails, you lose 1 x your investment; but if you invest into a project and it's a great success, you may make back 1,000 x your investment. So investors want to invest into many (say, 100) moonshot projects hoping that one will succeed.

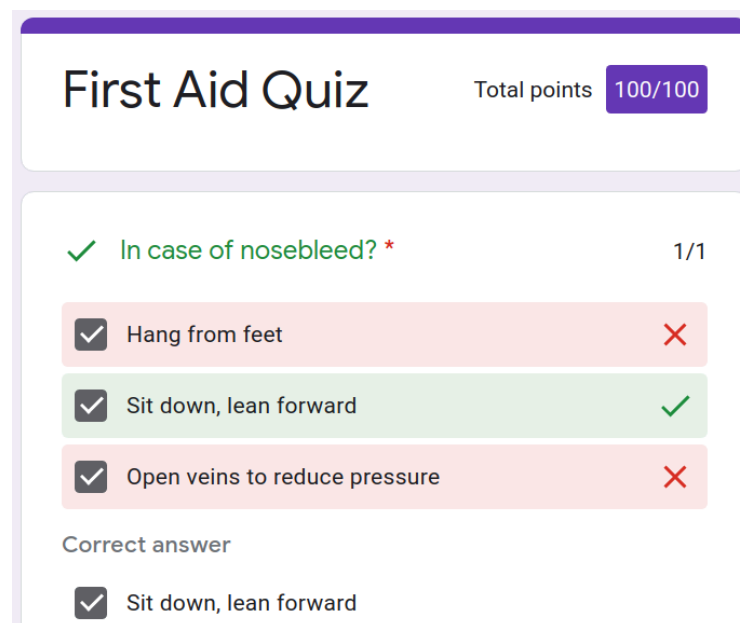
When it comes to for-profits, governments are to some extent trying to limit or tax externalities, and one could also argue that if one company didn't cause them, then another would've done so only briefly later. That's cold comfort to most people, but it's the status quo, so we would like to at least not make it worse.

Charities are more (even more) of a minefield because there is less competition, so it's harder to argue that anything anyone does would've been done anyway. But at least they don't have as much capital at their disposal. They have other motives than profit, so the externalities are not quite the same ones, but they too increase incarceration rates (Scared Straight), increase poverty (preventing contraception), reduce access to safe water (some Playpumps), maybe even exacerbate s-risks from multipolar AGI takeoffs (some AI labs), etc. These externalities will only get worse if we make them more profitable for venture capitalists to invest in.

We're most worried about charities that have extreme upsides and extreme downsides (say, intergalactic utopia vs. suffering catastrophe). Those are the ones that will be very interesting for profit-oriented investors because of their upsides and because they don't pay for the at least equally extreme downsides.

Most profit-oriented investors also care about countless things besides profit, but we think it makes sense to think about these risks with a security mindset and assume that there are purely profit-oriented investors out there who can move a lot of capital. Besides, most investors will not be aware of the interests

of future generations millions of years from now, of invertebrates, or of beings close to our acausal trade partners.



First Aid Quiz Total points 100/100

✓ In case of nosebleed? * 1/1

- ✓ Hang from feet ✗
- ✓ Sit down, lean forward ✓
- ✓ Open veins to reduce pressure ✗

Correct answer

- ✓ Sit down, lean forward

A simplified analogy is pictured above. This first aid quiz counts correct answers only – just like the market – so if you select all answers, you get the full points even though most people would not survive such treatment for nosebleed.

Anthropic distribution mismatch. Moreover, if the downsides are strictly about extinction, then the investors will lose their bets in worlds in which they wouldn't have been able to spend the money anyway.

They might regret this if they thought that their bet itself had increased the probability of losing the bet, but they'll probably assign a low probability to that because they may, for example, reason that they merely defected in an already hopeless collective prisoners' dilemma. Some people reason along the same lines when they argue that divestment is ineffective ([something that Paul Christiano critiques](#)): If the market is sufficiently efficient, any divestment will result in an inefficiency that will quickly be compensated by equally sized investments of others. These factors cause us to worry that investors will be likely to defect against values that are concerned with x-risks (including s-risks).

This could be framed as a problem of moral trade because there are those who care about the continued existence of our civilization and those who care about profits (and the continued existence of our civilization), and the first group may be ready to pay the second to divest from civilization-destroying charities in favor of other profitable investments. There are probably even better possible compromises like that.

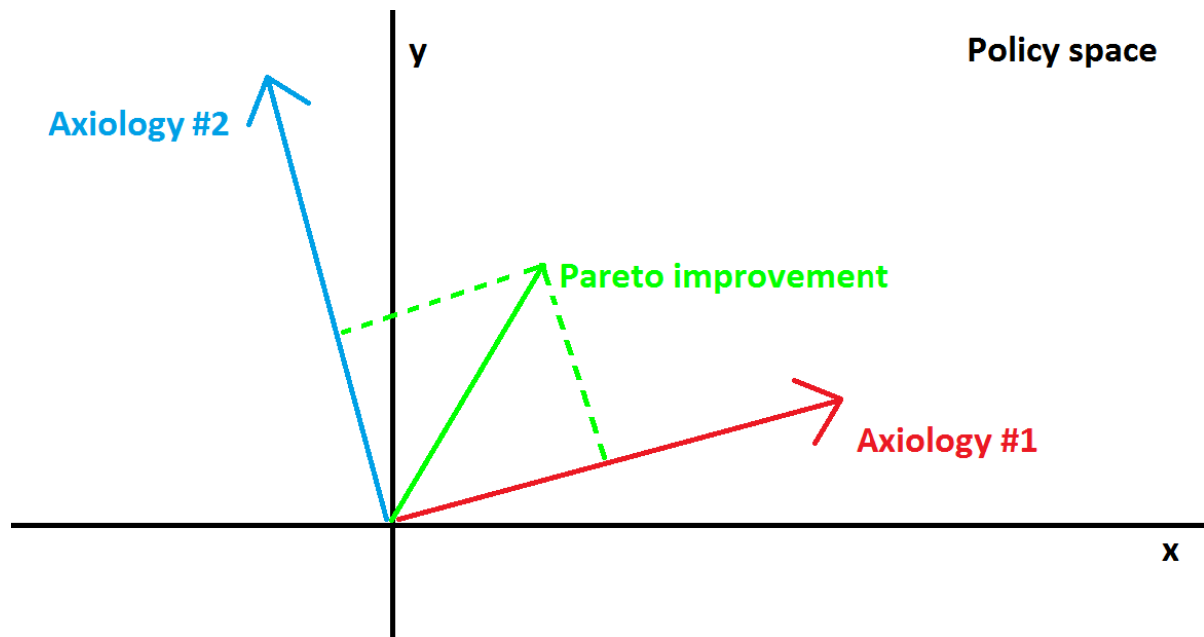
But that requires at least one of two things: That the for-profit investors take x-risks and their aversion against x-risks seriously, or that the investors who are mainly concerned with x-risks collectively have a lot of capital at their disposal.

Below we present a definition of "Attributed Impact" that combines earnest intention with outcome and thereby addresses the problems related to moral trade and extinction.

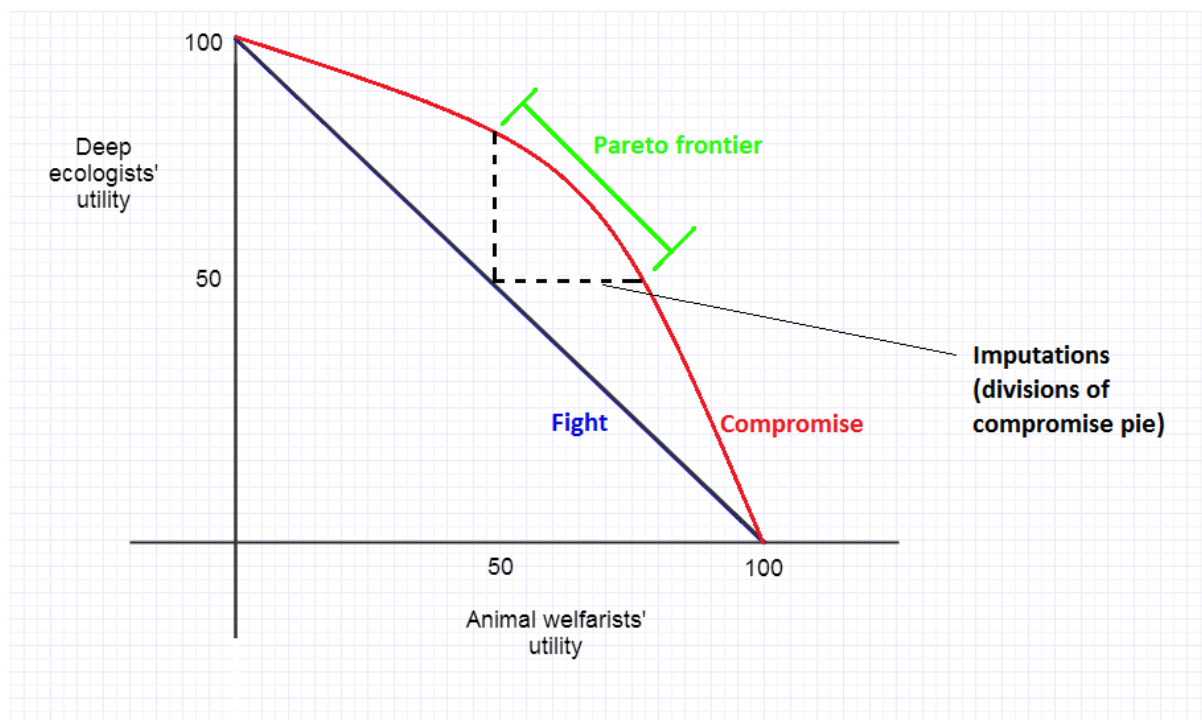
Moral Cooperation Failure

It is uncommon for two people to each have only one interest, and for these interests to be exact opposites. So in most cases it should be possible to find a compromise that improves the aggregate good that is achieved and, ideally, one that is as good or better than the cooperation failure for both, a Pareto-improvement.

Here are some pilfered images by Brian Tomasik from the [Center on Long-Term Risk](#):



The caption is: "Pareto improvements for competing value systems. The two axiologies are opposed on the x-axis dimension but agree on the y-axis dimension. Axiology #2 cares more about the y-axis dimension and so is willing to accept some loss on the x-axis dimension to compensate Axiology #1."



Caption: "Figure 2: Imputations for compromise between deep ecologists and animal welfarists, with $\pi_i = 0.5$ for both sides. By "Pareto frontier" in this context, I mean the set of possible Pareto-optimal Pareto improvements relative to the (50, 50) disagreement point."

The line labeled as "Fight" is the conflict that we would see on our markets if both parties were to fundraise for their conflicting goals and consequently burn most of their resources in a zero-sum conflict. But if they talk to each other and compromise, then they can realize outcomes that are better in aggregate (almost inevitably), and in particular outcomes on the Pareto frontier, which are directly desirable (or neutral) for both parties!

Had impact markets existed a few decades ago, there might've been projects that pushed for coal energy over nuclear energy. These might then have gotten a lot of funding from investors and retroactive funders who didn't think that climate change was a significant worry. Our and future generations would then have been defected against. (Our interests would've been ignored because we didn't participate in the market through retro funders.) Instead it would've been possible for these people to instead invest into R & D for the next generation of safer nuclear reactors, which would've allayed their safety concerns without exacerbating climate change.

Similarly, an animal conservation organization might want to use impact markets to fundraise for the protection of certain predator animals. They might fundraise large sums from people who are interested in the protection of that species, but they would thereby defect against the interests of the prey animals who are much more numerous and probably similarly capable of suffering. Instead they could've fundraised for the protection of a herbivore species that mostly eats fruit and whose members are unlikely to experience great amounts of suffering themselves throughout their lives.

Worse, the conservation charity that increases the number of predators at the expense of the prey animals may sell impact at a positive price and another charity that protects the prey animals against the predators by (say) making them infertile could also sell their impact at a positive price. Both impact certificates may

have a large positive valuation when really their impact more or less cancels out! So this egregious waste of resources can even happen when both parties participate in the market in some way.

Another egregious example is terrorism. Terrorists may want to use impact markets to fundraise for their attacks. Maybe they are religiously motivated and are attacking people as a form of proselytizing with extreme prejudice. Meanwhile other groups may use the same markets to fundraise for terrorism prevention. That is again wasteful since they could instead pool their resources in open-access adversarial collaborations on questions of religion.

We think this lack of any incentive for individuals to engage in moral trade is a major risk. The distribution mismatch is, strictly speaking, a sub-risk, but that's a bit unintuitive, and we think that both are so important that they deserve their own sections be it only for emphasis.

Finally note the difference between *revealed* and *idealized* preferences. [The first step](#) when attempting a compromise should probably not be to trade right away but, if possible, [to discuss the object-level disagreement](#). It may be that at least one of you is simply wrong, will realize that, and will no longer be interested in the competing intervention at all. (But others might be.)

Manipulation of the Default

An attacker might first build a reputation of regularly doing exactly the sorts of things that large funders hate, such as running flash loan attacks against crypto markets or denial of service attacks against blockchains, and then skip some of these regular attacks in order to sell their impact from *not* attacking. This exploit is based on the ability of the attacker to change the default state of the world from one in which they don't attack to one in which they do attack, and regularly.

This failure mode would cause impact markets to exacerbate *exactly* the sorts of problems we want to solve. The "Attributed Impact" definition below addresses it.

One real example that sort of fits is the following anecdote that I've heard from a reliable source but haven't fact-checked: When the cap & trade carbon trading system was first introduced in the EU, many polluting companies either first polluted even more or exaggerated their level of pollution so that their default level of pollution would be set higher. With the exaggerated default it was cheap for them to (seemingly) reduce their carbon emissions again, and they didn't have to buy carbon certificates.

Similarly, one [sometimes hears of stores](#) that momentarily increase the price for a product (or just lie about the old price) and then advertise the same or a higher price as a discounted sale price.

We're not as worried about this as we are about moral trade and, in particular, the distribution mismatch, because attackers could, in many cases, already attack funders this way. The funders just have less of a target painted onto them, figuratively speaking. Yet funders should make it clear that they will not be fooled this way, because even a failed attempt at such a ruse can do harm, as in the case of the increased pollution.

Drawbacks of Verifiability

Impact markets will probably tend toward high verifiability requirements all by themselves, but only after many buyers have been burned by investments into impact certificates that later turned out to overlap with

other impact certificates and whose price dropped as a result. Prescribing a high degree of clarity from the start will hopefully avoid this friction and keep more buyers interested in impact markets. Keeping impact certificates more clearly delineated also helps to separate them from impact stock.

But there are two possible drawbacks to this. The first is related to Goodharting. Charities that have a very clear output – e.g., papers or bed nets – will have an easy time fundraising on impact markets. But charities with more fuzzy outputs – e.g., community cohesion or plan changes – may be relatively disadvantaged by the format. That could lead to a relative underinvestment of money and effort into these more fuzzy interventions. Worse, most interventions can probably easily be measured by a few proxy measures that, when optimized for, lead to terrible outcomes. That's a case of a failure of moral trade. Attributed Impact discourages it, but overconfident issuers may still seek to sneak in the bad proxy measures that they perform excellently on.

The other drawback is that explicit contracts with collaborators may not be enough. We're currently collaborating with many others in a range of ways. Some examples in rough order of how explicit they appear to us: collaborations between coauthors, collaborations with reviewers, collaborations with casual conversation partners, collaborations with providers of infrastructure, and collaborations with those in the past who have made it possible that we are alive today.

If the sale of an impact certificate for, say, a paper rewards only the authors, there's a risk that other collaborators may decide that they will be rewarded better if they don't engage with other authors anymore to fully focus on their own papers and not give away any ideas before they have published them.

This could also happen between organizations. Each may think that it has a uniquely important mission and a responsibility toward its employees, and that it may have to shut down without funding from the sale of impact certificates. Two seemingly altruistic reasons will push these organizations to withhold all sorts of resources from the other.

The solution is to talk to all collaboration partners that form a coalition and to negotiate an allocation of the returns from impact certificates that has the core property, that is that no collaborator thinks that it is in their interest to split off from the coalition.

That's easily done among explicit collaborators such as coauthors, cofounders, or employees. Hence why the first rule requires explicit contracts between these. But in many other cases it will require acausal trade because it would be infeasible or impossible to (causally) negotiate with the collaboration partners – they may be anonymous, dead, distant, too many, or too busy. So the first rule falls short of solving this problem.

On the other hand, impact markets are likely to reduce financial scarcity, and people are less likely to behave uncooperatively when they have plentiful resources.

Noise

The market itself will also serve as a source of information for investors to understand what impact likely maximizes the compromise morality. Investors who correctly identified highly uncontroversially valuable impact early on may sell to take profit, manage their risk, or pay their rent – reasons other than thinking that the impact is overvalued – but this sell may be misinterpreted by others as a signal that the impact was more controversial than they had thought. All of this introduces a lot of noise that distracts from the price discovery. This is something that happens on financial markets all the time, and public companies

have to be sophisticated enough in their budgeting not to be ruined by fluctuations that have nothing to do with their fundamentals.

Sophisticated funders who are highly confident in their judgment, or more so than other market participants, can stabilize the prices by buying, holding, and dedicating impact certificates, but they can do so only to the extent that their budgets allow.

This is more of an inefficiency than a risk of net harm, so we're less worried about it than about the above.

Solutions

(We briefly covered this section in [our talk at Funding the Commons II](#).)

Attributed Impact

We think it will be key to find a definition of impact that has four properties:

1. It leads investors to value impact in proportion to the moral gains from trade that it has generated,
2. It leads investors to value impact in accordance with some form of idealized rather than revealed preferences of all affected groups,
3. It forms an attractor state so that over time more investors tend to adopt the definition (be it unconsciously) rather than fewer, and
4. It tracks but makes more precise the existing shared intuition of what "someone's impact" is.

The following definition is an attempt at that. It relies on a lot of culturally shared understanding, which is hopefully also shared to a sufficient degree between investors. Enforcing honesty or detecting lying of market participants is outside the scope of the definition. We'll call this impact *Attributed Impact* to avoid clashes with other common meanings of *impact*. (We're capitalizing the term not for emphasis but to clarify that it's not descriptive, i.e. just thinking of impact that is attributed to someone does not fully capture the definition.)

Definition

Attributed Impact (vo.2) is:

1. a responsibility, justified by those who make the claim,
2. for the minimum of the subjective expected change
3. in the aggregate supply of *all* public goods at
 - a. the moment when the issuers made the decision to generate the Attributed Impact and
 - b. this moment,
4. compared to the set of counterfactual world histories that are most unsurprising to an observer who knows nothing of the specifics (issuers, project, time frame, etc.) of this Attributed Impact,
5. where *subjective* refers to the epistemic perspective of each market participant, who should, by default, be assumed to
 - a. use a noncausal decision theory and
 - b. favor the Kalai bargaining solution.

In Plain Terms

As issuer of an impact certificate:

1. You need to justify that the actions you're planning to take will generate attributed impact. You need to reference the definition (including its version) and make the argument in writing in or attached to your impact certificate.
2. You cannot just pick one metric or moral dimension along which you want your impact to be measured. You need to make the argument that your action is good or neutral for all moral views, that harms are offset in ways the participants would accept, or something along those lines. You can also think of it as making an argument for the magnitude of the gains from moral trade that your action will generate. In short, *an' it harm none, do what ye will*. (Slight oversimplification.)
3. In particular, you need to argue that at this time (when you haven't taken the actions yet) it is reasonable for the median market participant to think that your action will generate impact in expectation. This should be positive impact unless you're happy that your impact certificate will have negative value and no one will buy it.
4. Over time, as you perform your action, the variance around your expected impact should shrink. It may then turn out that your impact is greater than expected. But the overall Attribute Impact is capped at the expectation going into the project; it cannot be greater. Ideally, you'll get it just right.
5. Note that you can't inflate your impact by claiming that something unusually terrible would've happened if you hadn't performed your action. If such a claim is not plausible to someone like John von Neumann, or anyone else who doesn't know and knows nothing about you, then it doesn't count.
6. Note also that you can't lie about your actions. If you write an impact certificate for one action and then then perform a different action, the impact certificate is for an action that didn't happen and there is no impact certificate for the action that you did do.
7. Finally, talk to others who you are bargaining with if at all possible. If not, please don't automatically assume that they make decisions according to CDT and if you're unsure about what bargaining solution they'll favor, go for Kalai.

Commentary

Responsibility and defense. Especially the term “responsibility” is one that relies on a culturally shared understanding, ideally one codified in law or a contract. We see no objective way to attribute impact other than by convention. Here the burden of proof is upon the person who makes the claim.

But they not only need to defend their legal right to the impact but also all other clauses of the definition. If every impact certificate thus recapitulates the definition, it will gradually become reified culturally as the consensus definition of Attributed Impact. We call this a “driver of adoption.” Attributed Impact contains several for redundancy.

Minimum of expectations. The minimum of *ex ante* (“the moment when the issuers made the decision to generate the Attributed Impact”) and *ex post* (“this moment”) expectation can be thought of as follows: To generate positive Attributed Impact, you need to earnestly intend to generate positive Attributed Impact. That is you cannot (1) generate it accidentally, (2) generate it while expecting that it might as well be vastly positive or negative, or (3) delude yourself into thinking that it will be positive through self-hypnosis or not doing any research.

And to generate positive Attributed Impact you also need to actually generate positive impact.

(We're referring only to *positive* Attributed Impact here since we don't see a reason why anyone would want to defend a claim to negative impact or buy someone else's negative impact.)

Thus Attributed Impact follows (in spirit) the [JTB definition of knowledge](#) as a *justified, true belief*. In the impact certificate or some ancillary document, the issuers justify their claim to the Attributed Impact by laying out how they think the reader ought to arrive at the conclusion that they were, at the time of the decision, justified to believe that what they did would generate positive impact. Time and the market then arbitrate the question of truth. Only if *ex ante* and *ex post* impact are positive is the aggregate impact (the minimum) positive.

This avoids the pathological case of *x-/s-risk gambles* where someone runs a large number of very risky projects – projects that can turn out vastly net positive or net negative – and then makes money by selling only the impact from those that happened to turn out positive even if they are in the minority. (Or likewise for funders.) Those that happened to turn out positive would still not be worth anything because their *ex ante* or negative *ex ante* impact will be less than the positive *ex post* impact, so that the positive impact will have no influence on the eventual value. This mirrors the intuition between expectational consequentialism where, for example, a doctor who treats a mild headache with a medicine that is 99% likely to kill the patient acts immorally even if the patient actually happens to survive and be cured of the mild headache.

Note that this applies in the same way to the anthropic formulation of the *x-/s-risk* gamble where issuers or investors believe that (1) they hardly affect the odds of extinction, for example because of market efficiency, and (2) they are neutral about *x-/s-risk* futures because they can't spend money anyway if they're dead or otherwise incapacitated but they do care about the okay outcomes where they can spend money.

Aside. The similarity to the JTB definition suggests that the [Gettier problem](#) may be a problem. But we haven't managed to construct an actual pathological or exploitable version of it. (You can skip to the next subsection if you're in a hurry.)

An example: Aanya Altruist reviews some 450+ studies and concludes that there is strong evidence that the serotonin transporter gene 5-HTTLPR is associated with depression. She develops and promotes a school-based screening procedure that provides people with the particular versions of 5-HTTLPR with resources on how to access therapy. Biology teachers love the program for how helpful and scientific it is. Aanya makes a case for her impact and sells it. Then it turns out that (1) [5-HTTLPR has nothing to do with depression](#), and (2) many of the children who had received the resources had become depressed at the population base rate but started therapy at a much higher rate.

Aanya Altruist's defense of her impact is state of the art by the standards of her field at the time, and she is honest about believing her conclusions. Yet they turn out to be false. So her impact is valuable by dint of her earnest, justified intentions and by dint of her quite unrelated impact. This is probably not going to be a common case, because we expect there to be few impactful interventions so that few people will find them by accident, but it doesn't seem harmful either.

But the system is hypothetically exploitable in another way: Someone may find a new, previously unknown CO₂ sink that they predict to start absorbing CO₂ within a few years. They build a complex contraption that no one understands but that is just plausible enough that they can fool investors with an amphigorical justification. Then they start it right when the natural CO₂ sink goes into action. Finally they claim credit for the impact.

Another variation of the theme: Someone builds in secret a lot of machines with fairly general seeming purpose but without actual function. Then they wait for a near catastrophes. Maybe there's a hurricane that changed course just before it hit the shore. They reveal

one of the machines that happened to be nearby and claim to have used the hurricane as a beta test for the hurricane-averting machine and sell the impact from it.

Maybe people will come up with more realistic variations on this theme. When that happens, the problem can probably be addressed through preregistration of impact certificates and buyers penalizing any lack of preregistration. (Preregistration may be as simple as issuing an impact certificate before the impact has happened, which should be the norm anyway.)

Subjective expected change. Refers to the expected value of the impact from the vantage point of a market participant, not from the vantage point of the issuer. The issuers may be incentivized to lie about what they knew at the time, but it doesn't matter if sufficiently many market participants think that the issuers ought to have known it at the time.

People tend to be biased to think that they would've already realized something earlier if they had considered it or even to [misremember their past predictions](#) as being more right than they really were.

There is a risk that people will default to assuming that other market participants are causal decision theorists. This would preclude gain from acausal moral trade. We're unsure what decision theory is best to recommend here, but it is probably not causal decision theory. The bargaining solution is also something we're unsure about. We don't know which bargaining solution to pick here, [but not specifying any seems worse than specifying a random one](#). Resource monotonicity seems to us like an important criterion for the general acceptability of a bargaining solution in a rapidly growing market (the Nash bargaining solution doesn't ensure it), maybe more so than scale invariance (the weak point of the Kalai one). Independence of irrelevant alternatives seems superficially important to us (the Kalai-Smorodinsky solution doesn't ensure it), but we haven't thought about this in detail and would greatly appreciate feedback on this decision.

Aggregate supply of all public goods. The "aggregate supply of all public goods" is an attempt to codify that what makes an action valuable are the gains from moral trade that it generates. (And potentially voluntary sacrifices of private goods to turn them into public goods.) The "all" in the phrase may be overly ambitious since there is no market mechanism to enforce respect for public goods that none of the market participants know about, but it probably doesn't hurt to aim high.

This exclusive focus on the gains from moral trade has a host of related advantages: Investments into zero-sum games will have no value on a certificate market, except insofar as they generate net positive externalities; progress on some values that is offset by harm to other values is without value; projects are incentivized to seek out new moral dimensions orthogonal to all known ones; and projects are encouraged to do cross-cutting work that benefits a variety of moral systems.

Set of counterfactual world histories. The "set of counterfactual world histories" is another point where we can only appeal to shared intuitions, since, in deterministic worlds, no other world history is possible. Note that we always compare full world histories, so not only the time between the decision to implement the project and the present moment or any other such period. This is written in the spirit of [a belief that the Tickle Defense is probably true](#). So a person who surprisingly kills someone produces evidence that people sufficiently like them have killed people before and thus narrows the set of subjectively plausible pasts to a set of worse pasts, but a historian who researches past killings has not the same evidential effect because

the killings they research screen off any evidential effect the historian has. The spirit of our definition is that if the Tickle Defense fails, we want to correct our definition and not put blame on historians!

The sets of world histories all contain several pasts, presents, and futures, because the issuer had uncertainty over past, present, and future at the time of the decision. (In view of the many-worlds interpretation, we intend “world histories” to include all Everett branches weighed by their measures.)

An observer who knows nothing of the specifics. This clause prevents issuers from manipulating the default as described above.

According to the above definition, changes in the supplies of public goods need to be assessed against the counterfactual that is the perspective of an observer that knows nothing about the specifics of the case. We'll call this perspective the *systemic* stance, in reference to Dennett's [intentional stance](#) and system science, which is a perspective that abstracts from agents and intentions to recognize systemic or emergent mechanisms. It's a form of underfitting or dimensionality reduction. As a result, the attacker would now have to change large swaths of society into one that regularly conducts flash loan attacks before their omissions to do so becomes different from the default.

Retro funders need to proactively signal their adherence to this definition of Attributed Impact to make it clear to any potential attackers that they are not vulnerable to this exploit.

Summary. This definition of Attributed Impact addresses (hopefully successfully) the following problems:

1. **Impact is not naturally uniquely owned.** It addresses this by relying on cultural norms and laws around responsibility, such as legal ownership, authorship, etc.
2. **Issuers may generate one public good at the expense of another.** This is addressed by tying the evaluation to an aggregate of *all* public goods.
3. **Markets incentivize x-/s-risk gambles.** It addresses this by defining Attributed Impact as the minimum of *ex ante* and *ex post* expected impact.
4. **Counterfactuals are impossible and arbitrary.** This is addressed by stipulating a particular way of thinking about the counterfactuals that relies on the subjective expectation at two points in time of an observer that knows no specifics about the case.
5. **Buyers are vulnerable to extortion.** This is addressed by stipulating a counterfactual (see above) that is very hard for an individual to influence.
6. **This definition may be ignored.** It addresses this by asking issuers to defend their impact in terms of the definition and by protecting buyers from extortion if they proactively signal adherence to the definition.

Timeline



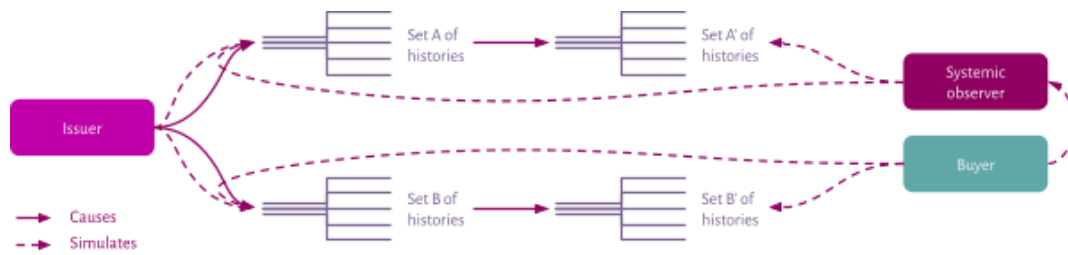
This is a sample timeline of how an issuer may produce Attributed Impact. Period 1 is when the impact is issued and generated, in whatever order. Period 1 separates Period 0, where the real and the counterfactual worlds haven't diverged, from Period 2, where they have diverged and continue to diverge.

On both sides of Period 1 there is one systemic observer, who is imagined/simulated by the investors. In Period 2 there is also the actual investor. (We'll use *simulate* over *imagine* to convey that the goal of the mental simulation is accurate prediction and not, say, entertainment.)

So this is what happens in temporal order:

1. The prospective issuers consider that a systemic observer would typically expect people like them to do *stuff*. So they surmise that future buyers would also agree that a systemic observer would expect that. But if they did a *thing* instead, a systemic observer would be surprised.
2. The prospective issuers iterate through a few options for *things* to do, and eventually settle on one. That *thing* is a great *thing* because it generates morally relevant preference satisfaction for some beings without undermining the preference satisfaction of others. They implement it and document it in an impact certificate alongside a lengthy defense.
3. A month has passed and a few buyers get interested in the impact certificate.
 - a. They by and large agree with the issuers on the expectation of the systemic observer in Period 0. This systemic observer is virtually the same as the one in Period 2. They imagine themselves in the shoes of the issuers in Period 0, and find that they might have predicted a few additional benefits and drawbacks but nothing substantial, and agree with those that the issuers have documented in the impact certificate, and with their conclusion that the *thing* appears like it would be net positive in expectation.
 - b. Then they return into their own shoes where they have the invaluable benefit of hindsight. From this vantage point they realize that a few of the drawbacks they were retroactively worried about had not manifested but that some of the best-case scenarios have also become less likely. All in all, the thing still appears to be net positive in expectation, now with slightly reduced variance. They can't quite tell which expectation, *ex ante* or *ex post*, is lower, but they think that both are positive and fairly close together, so they don't care much and just bid on the impact certificate according to their budgets.

Note that strictly speaking the buyers also simulate the issuers because the certificate will be phrased in natural language which is highly ambiguous without knowledge of the (likely) intent of the author. We'll omit this simulation in the graphs since it's quite intuitive.



That's all a bit of a simplification because many prospective buyers (let's call them investors) want to make a profit. So they'll perform further evaluations to understand whether they may know more or have more accurate information about the real expected impact than other investors and whether the other investors will attain that knowledge too, be it automatically over time or through a publication by other investors. Occasionally they may also seek to prove that the issuer had private information already in Period 0, which they pretended not to have, or conversely, whether something that is widely assumed to have been known at the time actually wasn't knowable.

Example

Example. The issuer has sold 500 pins with cute designs and animal rights messages on them at a convention. They've donated the proceeds to an animal rights charity. They now sell the impact from this action.

1. The issuer thinks that the donation of the proceeds is the main driver of the positive impact, so the certificate text focuses on the money.
2. But one investor thinks that most investors are unexcited by the donation because they think it just displaces donations from corporate partners of the charity who don't have set CSR budgets.
3. This one investor thinks that most investors think that reaching 500 people with the animal rights messaging is the main driver of the impact, in particular because the designs were actually created by well-known members of the community that the issuer sold the pins to, something that is widely known but that the issuer didn't mention.
4. But this one investor also freshly found photos from the convention and noticed that the issuer had sold five different designs, and that, in group shots, a lot of people could be seen wearing multiple or even all five pins. The investor shorts the certificate before announcing their finding because they expect that the issuer had reached much fewer than 500 people.

This example aims to clarify that:

1. Issuers and investors can disagree on the path to impact, yet it can still be valuable *ex ante* and *ex post*.
2. The certificate description is never going to describe reality fully (and critical omissions can be unintentional), so it should contain as much information as possible that will help buyers do their own research to fill in gaps – such as finding photos from the convention.
3. The issuers and all investors can have different ideas of the actual and the counterfactual world histories from the perspective of the systemic observer.

4. Only when an investor can anticipate a price-relevant update of some of the other investors can they make money. If an investor knows more than the others but has no way of convincing them, they can't scoop them. (But in this case they could because they had photographic evidence.)

Responsible Retroactive Funders

The main responsibility for the health of the market is upon the retro funders. Hence they need to be sufficiently wise to steer the market and sufficiently involved to notice when something is going wrong in the market.

First, a retro funder needs to ask themselves:

1. How much time are we spending on finding giving opportunities whose interventions are effective and that are implemented by capable teams? How much is it worth to us to not have to evaluate the teams? How much are we willing to offer to investors to incentivize them to do the work for us?
2. Do we think that in the future people will value the impact that we're interested in even higher, so that we can one day resell it?
3. Do we want to see more highly effective charities, and how much are we willing to pay to incentivize their founding?

If these questions generally turn out that the retro funding will be worth it, there are further consideration:

1. Investors and founders will rely on our future funding (not that they'll get it necessarily but that it'll still be there). Can we commit a certain budget firmly, and if we need to discontinue it, can we afford a few years of grace period between the announcement and the discontinuation?
2. Can we make it sufficiently clear to founders and investors what the impact is that we want to see, so as to prevent people from expending great effort on potentially harmful projects that we would never pay for?
3. Are there maybe even metrics that can be used as rough guides by founders to check whether they're on the right path without great risks of Goodharting or moral defection?
4. Do we have the resources for occasional calls with founders or investors who want to be sure that they're starting the right sort of project?
5. Do we have resources to occasionally write a blog post to clarify our interests when we see investments being poured into uninteresting projects?

Finally the hardest question:

1. Are we ready to commit to being as impartial as possible about our impact evaluations? Especially in cases where we have strong moral feelings in one direction or the other, can we be trusted to still reward only that project that successfully compromises between the competing interests? Or put differently: Can we be trusted to price projects in proportion to the moral gains from trade that they generate?

If someone is only interested in providing retro funding but not in doing all the work associated with being a responsible retro funder, then the next section will be interesting, as they can simply contribute to the "Pot of Money" and leave the work to the jury of the pot.

Pot of Money

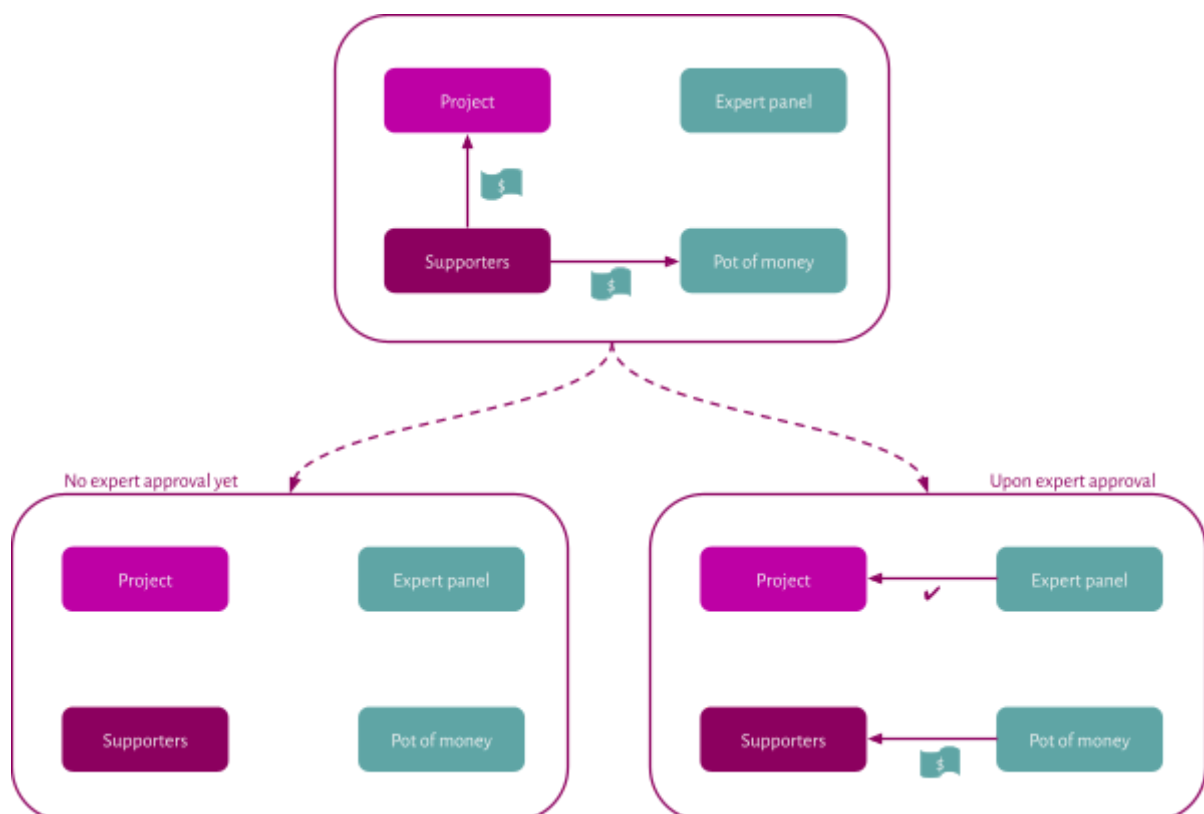
If there are sophisticated, responsible retro funders on a market and if they have a lot of money to give away compared to the volume that is transacted on the whole market, then everything is well and good. But even if either (1) there is no retro funder, (2) there are sophisticated, responsible retro funders but they are overpowered by rogue ones, or (3) there are only sophisticated, responsible retro funders but speculation on the market has developed its own rules, then the influence of the good retro funders will be insufficient to keep the market healthy. This is where the pot comes in.

The pot is a reference to the pot in poker, a vehicle for investors to “gamble” – except that no intentional randomness is involved, only the imperfections of impact evaluations and of predicting the future. For simplicity, we’ll copy the example from earlier but rewrite it to the case where the only retro funder on the market is the pot.

There are now four components:

1. **Projects:** The charitable projects that are waiting to be realized.
2. **Supporters:** These include founders, investors, advisors, et al.
3. **Jury:** An expert panel of sophisticated altruists, who become the source of truth when it comes to what has been impactful.
4. **Pot of money:** When the sophisticated altruists publish their decree, this pot of money is dispensed to the supporters.

At first, the pot of money likely has to be managed separately. The charities are not a good fit for managing it, and the funds may not be sold on the whole system from the start, or not sufficiently for all the overhead it would involve for them.



Let's suppose again that the Against Malaria Foundation (AMF) wants to do a distribution of long-lasting insecticide-treated bednets with its old distribution partner Concern Universal (CU):

1. AMF ascertains that the jury of the pot is (and maybe other retroactive funders are) still interested in more net distributions.
2. The AMF staff decide how they want to split the Attributed Impact from the distribution among themselves, the legal entity AMF, all their partners, contributors, and advisors, and their funders. These decisions can be adjusted in later negotiations. (we'll treat AMF, the legal entity, and all its staff as just "AMF" in the following to save space.)
3. AMF decides on the details of the auction process, that is, what fraction of the profits from each sale should go to AMF and what the minimum percentage raise is between bids.
4. AMF creates an impact certificate.
5. AMF contacts CU, negotiates their split with them, and sends them their fraction of the impact certificate. (we'll assume here that the distribution partner has its own source of seed funding just to make its case interestingly different from AMF's.)
6. AMF contacts a venture capital firm. The VC agrees to fund the distribution. They negotiate the split, and AMF sends them their fraction of the impact certificate.
7. Everyone contributes to the pot as they see fit (or doesn't), and the pot records who contributed how much to bet on which certificate.
8. A year later the distribution gets completed successfully.
9. The jury notices the successful distribution, is ecstatic, and buys fractions of the certificate from all holders who have contributed to it. AMF has no work with this: Any holder just has to give their certificate to the pot (or a fraction thereof), which ascertains that they've previously bet on it, and the pot buys it from them at a higher price.

The core of the system should, in our opinion, not depend on any additional retroactive funders because (1) they may not be interested in the market until they see it work; (2) their funding is finite and independent of the impact market; and (3) they may not have enough in-house experts so that they want to defer to our jury anyway, effectively adding to the pot.

That said, external retroactive funders would make the market attractive for many more participants!

When it comes to the size of the reward, we currently favor the following system:

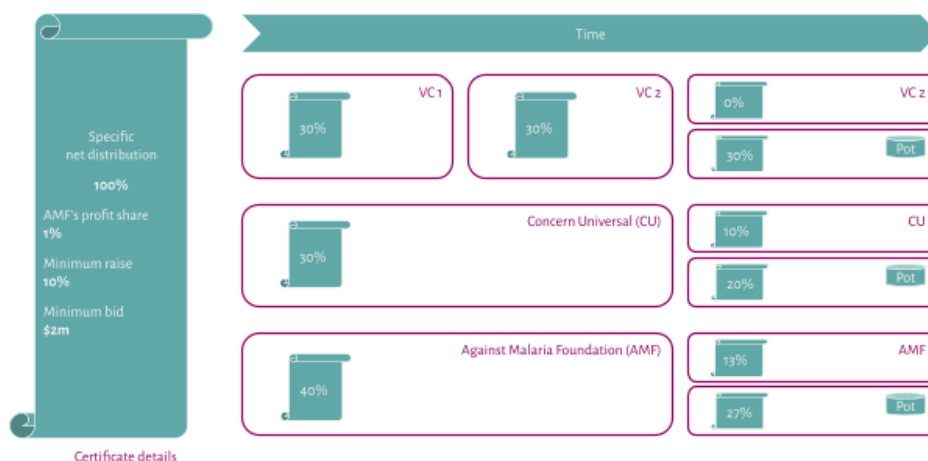
1. Everyone contributes whatever they want but has to tie this "bet" to a particular impact certificate.
 - a. This removes incentives that investors who don't like the pot might otherwise have to circumvent the market.
 - b. They need to bet on a particular certificate rather than just any that they own so that they can't just buy minimal fractions of countless certificates at random, which would add little to no wisdom to the market and wouldn't differentially benefit projects with real promise.
2. When the jury approves of a certificate, the pot sets a budget for the purchase and buys fractions of the certificate from each buyer according to the product of (how much of the certificate they hold) * (their contribution to the pot for that certificate).
 - a. That way people are rewarded for contributing to the pot and the project, and those who fail to do either can't sell to the pot.
 - b. They may of course do so on purpose because they want to keep the certificate as a long-term investment or to sell it to another retroactive funder.

This still glosses over a lot of details. They are easier to explain in the context of the auction platform.

Auction Platform

We want to create an auction platform to streamline the whole process. ([You can keep up-to-date on the progress here](#).) The current model is a type of auction where the impact certificate is always owned by the highest bidder and so constantly changes hands. Meanwhile the profits from each sale are split between the seller and other parties, typically the issuer and the Pot. We've covered it earlier in the subsection Auction Platform.

In this model the issuer determines what fraction of the profits of each sale should go to them. Here's the rough timeline of the trades.



Here an overview of the accounts of each participant at each step in the process ([you can find the spreadsheet here](#)):

		VC1		VC2		CU		AMF		Pot	
Time step	Valuation	Share	Cash	Share	Cash	Share	Cash	Share	Cash	Share	Cash
Initial setting	n/a	n/a	\$1.0m	n/a	\$0k	n/a	\$500k	n/a	\$0	n/a	\$10m
Minting	\$2m	0%	\$1.0m	0%	\$0k	0%	\$500k	100%	\$0	0%	\$10m
Deal w/ CU	\$2m	0%	\$1.0m	0%	\$0k	30%	\$500k	70%	\$0	0%	\$10m
Deal w/ VC1	\$2m	30%	\$400k	0%	\$0k	30%	\$500k	40%	\$600k	0%	\$10m
Deal b/n VCs	\$2.2m	0%	\$1.05m	30%	-\$660k	30%	\$500k	40%	\$607k	0%	\$10m
Pot contrib.	\$2.2m	0%	\$1.05m	30%	-\$734k	30%	\$450k	40%	\$557k	0%	\$10.2m
Dist. complete	\$2.2m	0%	\$1.05m	30%	-\$734k	30%	-\$50k	40%	\$57k	0%	\$10.2m
Pot retro. buy	\$2.6m	0%	\$1.05m	0.1%	\$48k	10%	\$478k	13%	\$781k	77%	\$8.1m

Various parameters:

AMF profit share	1%		Pot frac.	20%		AMF cost	\$500k
Minimum raise	10%		Pot raise	20%		CU cost	\$500k
Minimum bid	\$2m						

In this scenario AMF made its profit share relatively small (1%) to encourage investment. The minimum raise (the step between bids) is at an arbitrary value of 10%. The minimum bid is set such that AMF won't raise amounts of money that are insufficient to fund the net distribution. The nets and salaries on AMF's side will be \$500k, and the salaries and miscellaneous costs on CU's side will be \$500k as well. (Traditionally, AMF's strategy has been to fundraise for the nets but to leave their distribution to other organizations.)

The deal with CU is described above. It's now that 30% of the certificate are issued. The deal with VC 1 is also described above. This is the first bid on the market, so VC 1 gets to buy them at the minimum bid price. VC 2 buys the shares from VC 1 right away, but this time the minimum raise and the profit share apply, that is, the valuation increases by 10%, and 1% of VC 1's profit goes to AMF instead.

Next, all three certificate shareholders contribute to the pot. AMF and CU add an arbitrary \$50k to the pot whereas VC 2 tries hard to maximize the weight it'll have when the pot buys the certificate from them again, but without selling more of the certificate than it owns. The prices in the ratchet auction can only go up, so even if it could borrow certificate shares from (say) AMF, that would not be profitable.

We assume here that (1) VC 2 is certain that the pot will later buy the certificate, (2) can wait for and then observe the pot contributions of all other shareholders, and (3) knows how much money is going to be in the pot and how much it will budget for this purchase. These are all slightly tricky assumptions: Assumption 1 is never going to be completely true; assumption 2 can lead everyone to wait until the last possible second to make their contributions; and assumption 3 is probably again somewhat difficult in practice.

If the jury were to deem the net distribution to have been ineffective, the scenario would end here with an extra \$174k in the pot and all organizations having lost money.

	Share	Value	Contrib.	Product	Weight	Cost	Size
VC 1	0%	\$0k	\$0k	0.00E+00	0.00	\$0	0.0%
VC 2	30%	\$792k	\$74k	5.86E+10	0.39	\$790k	29.9%
CU	30%	\$792k	\$50k	3.96E+10	0.26	\$534k	20.2%
AMF	40%	\$1.1m	\$50k	5.28E+10	0.35	\$711k	26.9%
	100%	\$2.6m	\$174k	1.51E+11	1.00	\$2.0m	77.1%

Finally, though, the pot decides on a budget of 20% and a bid 20% above the last one. The resulting valuation is more than the budget, so that the budget doesn't need to be capped. The weight is proportional to the value that the shareholder has in the certificate times the size of the pot contribution. ([You can find the spreadsheet here.](#))

The column "Share" is the share in the impact certificate that the respective party owns. The "Value" is the value of the share according to the latest valuation of the certificate. The "Contrib. [-ution]" is the amount

that the party has freely chosen to contribute to the pot. The “Product” is the product of the value of the share and the contribution – it doesn’t map to anything in particular in the real world, but the weight is proportional to it. The aforementioned “Weight” is the fraction of the budget of the pot that the pot uses to buy certificates from the respective party. The “Cost” is the resulting total dollar volume of the purchase. The “Size” is the fraction of the certificate that changes hands.

The final budget for each individual purchase may not be fully used up if any of the shareholders doesn’t own enough shares (though this case doesn’t happen here and is not part of the calculations in the spreadsheet because it made it harder for us to find errors in it).

The result is that the pot spent about \$1.86 million and every organization profited, either monetarily like the VC, altruistically like CU, or both like AMF.

Notes:

1. It took us a while to come up with a scenario that “works,” in the sense that in particular the VCs make a profit (or else they wouldn’t invest) without making the pot or the fraction that it invests unrealistically big or its bid unrealistically high. This indicates that the parameters of the pot budget will need a lot of tuning and that issuers and investors will be well-advised to carefully think through their funding scenarios to make sure that they’re realistic.
2. This difficulty is exacerbated since the pot will probably require holders to have held the certificates about a year earlier already to prevent frontrunning of pot purchases. Investors who hold certificates for a year will need to make much greater profits off them to make up for the counterfactual uses of their funds. Unless another investor buys it from them, and they buy it back much later.
 - a. The current auction system doesn’t support or would make it at all lucrative, but under different conditions, holders could lend their certificates to short-sellers to earn interest on them.
3. But this may be a feature rather than a bug in the short run: The distribution mismatch problem below is alleviated if (1) Attributed Impact has time to be adopted as the Schelling point of impact evaluation on the market, and (2) there are few purely profit-oriented investors. We may grow a safer, more altruistic community if the market is not of interest to anyone unless the social bottom line also counts into their personal profits.
4. There’s a questionable dynamic where some investors benefit if other investors into the same certificate pay less into the pot. We don’t yet know whether there are setups where it’s profitable for all investors if one of them pays other investors outside the market in exchange for them withholding money from the pot. But if there are such scenarios, then that may be bad for the pot.
5. There’s another bad dynamic, maybe a negative externality, where sophisticated for-profit investors are incentivized to advertise the system to unsophisticated investors and potentially mislead them. They may promote the impact certificates of highly ineffective charities to unsuspecting people as investment vehicles to use in this system. These people would enlarge the pot without having any chance of ever winning it. Meanwhile all the misinformation may make it harder to find high-quality charity reviews for people who don’t already know where to find them.
6. Finally, we’re unsure how to prevent insider-trading by jury members, but there are probably already mechanisms for that, such as that jury members need to be many and not know each other. There are probably time-tested solutions to this problem that we just need to find out about.

Prediction markets. Something we want to think about more is whether the Pot should instead be a scalar prediction market predicting something like the percentage of endorsement of the project from the jury. Investors may be hesitant to pay into the Pot if it's unclear for how long their money will be locked up in there even if they win their bet – a prediction market would address that.

Auditors

Our ideas for how to structure the ecosystem of auditors are still inchoate.

Auditors will need to audit impact at least twice. First there needs to be an audit that issuers need to get before they fundraise from investors. Instead of having to plow through all the impact certificates on the market, investors can then just consider ones that have been audited in this way. Such an audit needs to consist of a few basic checks, such as whether the argument for Attributed Impact makes structural sense (e.g., is not autogenerated gobbledy-gook), is concrete enough that it can be audited at all, and crucially that the impact is not being double-issued.

The second audit needs to also ascertain that the impact has happened. A lot of forms of impact are easily verifiable – an article author, for example, can just link to the article they've written.

Checking that impact has not been double-issued is hard in full generality, but it can be mostly solved if all auditors publish all their audits, and all auditors know about all other auditors and check their lists of published audits. Then they can't prove that impact has not been double-issued at all, but if it has, at least the other issue does not have an audit. This can still break down in cases where the double-issue is sufficiently subtle, but it should catch a lot of more obvious cases.

One thing that would be helpful for this job is a standardized format that all auditors use to publish their audits. Maybe a JSON blob on their websites. Then a search engine can extract structured data from there, and the search will be much easier for all of them.

However, we haven't thought enough about how to prevent auditors from colluding with nefarious issuers, because they are paid by issuers.

Targeted Marketing

Especially in the early days when the market is still young it'll be important who to market it to. The above section on responsible retroactive funders has already clarified why it is critical to recruit the sorts of funders who will let the markets thrive rather than ones that'll lead them astray. That is one problem that can be partially addressed by reaching out specifically to the retro funders who we think will do a good job of it.

But the same is true of charities and investors, just to a lesser extent. The charities on the market need to be the ones that produce good/s that is/are interesting to the sorts of retro funders we want to have on our markets. The reason is different in the case of the investors: If we recruit investors whose motivations are mostly altruistic, and we then notice that our markets create bad incentives for them, then we can warn them of these incentives, and it'll be in their altruistic interest to resist them. But if we run into bad incentives and have unaltruistic investors, we don't have this option.

Curation

A variation on the auditing that can be used in conjunction is curation. If an issuer doesn't have a large following themselves who they can market their impact to, they may want to rely on popular marketplaces to do so. If these marketplaces are curated by people with the same sophistication as retro funders, they can catch impact certificates that are too likely to be harmful early in the process, prevent them from being listed, and, if applicable, warn the issuer of the issues with the certificate.

Shorting

All of our solutions try to make it uninteresting to issue or to invest into impact that is likely net negative, but another option is to require investors to deposit collateral as insurance against the case where the impact turns out negative. The catastrophes that we're worried about are of the civilization-ending type, so no amount of collateral can realistically be enough, but it may still increase the friction a bit when it comes to investing into potentially net negative impact.

One idea is that we expect that in the long run most people will be agnostic about which impact certificate (e.g., which net distribution of AMF) they invest in because they're not close enough to the specifics of the intervention to see the differences between them. They'll just want to invest in "AMF." So I find it plausible that there will be demand for [perpetual futures](#) (see also "[What are perpetual swaps?](#)") that track something like the market capitalization of all impact certificates issued by AMF. (That's not an interesting market in the case of the ratchet auction where prices can only go up, but it may be with other auction mechanisms.)

To get exposure to this perpetual future, investors could deposit collateral and open long positions on it. If they use too much leverage and the price decreases too much, they can lose their collateral.

But the other benefit – hinted at in the subsection title – is shorting. Shorting can help to price in the expectation of some investors that retro funders will not be interested in the impact or not at that price. It doesn't by itself have the same upside potential as longing, but if a short-seller regularly reinvests their collateral when it increases due to decreasing prices, they can maintain a constant leverage and gain a similar exposure to nonanthropic downsides – at least before adjusting for the risks.

This sounds complicated, but there are already "hedge" tokens that automate the process. These could be applied to the perpetual future and thus give investors a simple, safe, and low-effort way to get negative exposure to overpriced impact.

However, we've seen these tokens fail especially during crashes, when you want to profit from them most. My hypothesis is that there is so little buy-side liquidity during crashes that when the token smart contract tries to reinvest its PnL to increase the leverage, it has very few buyers to sell to and so has to sell at low prices, prices close to where the market is crashing to anyway. They probably work better in markets where the prices decrease gradually.

All in all, it's much too early for this system. There'll first need to be highly liquid markets and popular issuers of many impact certificates before it makes sense to think about this system more.

Current Work

Protocol Labs has organized two iterations of the conference [Funding the Commons](#). “Funding Public Goods – Algorithms and Mechanisms” by Vitalik Buterin and “S-Process Funding” by Andrew Critch are particularly interesting because they highlight important mechanisms, and “Quadratic Funding on Gitcoin” by Kevin Owocki, “Retroactive Public Goods Funding Experiment 1” by Karl Floersch, and “Impact Evaluators” by Evan Miyazono are particularly interesting because they give insight into the work of the actors who are currently active in the space. Finally, “Impact Certificates and Impact Markets” by Owen Cotton-Barratt is the talk that we collaborated on and that we already linked above. [We also had a talk at the second Funding the Commons conference](#). We’re planning to summarize some of the talks in a separate article.

These talks highlight our work, the work of the Ethereum Foundation, the Survival and Flourishing Fund, Gitcoin, Optimism, Protocol Labs, and many more. Beyond that, Sentience Research, the EA Hotel, and Giveth have been following the field with interest.

We’re also in touch with [NPX Advisors](#) who’ve been running a quasi–impact market with greater amounts of funding but a very small set of funders, investors, and charities in the US, and who conceive of “impact certificates” as debt securities rather than impact equity securities.

We think these are viable approaches because they test impact markets in fields that are safer than some of the main EA cause areas. That way they either work with participants who are highly informed when it comes to financial markets or within fields where it is hard for us to see how impact markets may backfire catastrophically.

Note that we – Kenny Bambridge, Matt Brooks, Dony Christie, and I – can use funding: \$100k would enable us to invest more time into this project (about 70% more in my case) and any amount around \$10k to \$1m would allow us to conduct retro-funding experiments. This money could be merely committed to these experiments and wouldn’t need to be transferred in advance. We would then make grant recommendations according to guidelines that we would like to publish in advance in order to make the recommendations as predictable as possible. Please get in touch if you’d like to help with funding, advice, or otherwise.

Acknowledgements

Many thanks to my collaborators Kenny Bambridge, Matt Brooks, and Dony Christie! Thanks to Katja Grace and Paul Christiano for the initial inspiration. Thanks also to Owen Cotton-Barratt, Ofer Givoli, Jay, mqp, bryjnar, makoya, and Justin Shovelain for discussions and ideas. You’ll each receive 1% of the impact of this article. Thanks to Protocol Labs for the Funding the Commons conferences.