

Some updates to my thinking in light of the FTX collapse

Owen Cotton-Barratt, Q1 2023 (largely drafted in January, tidied in early March,)

Meta

What happened with FTX was a huge deal. It seems important to make a good attempt to learn all appropriate lessons from this. I think that that means talking about our updates a lot more. This doc will walk through my own updates at a level of granularity that I don't normally manage, with the hope that some of the thinking is useful for others. This writeup is based on a lot of thinking I did primarily in December and January, and the draft was largely complete by the end of January.

I want to acknowledge the awkwardness about me writing anything about EA community dynamics at the moment. In addition to the importance of giving people space to think through how they want to orient to [recent news about my past conduct](#), to the extent that people have questions about my judgement in some areas, they might reasonably also question it in others. I really don't want to be treated as an epistemic authority here.

I decided to share this draft more widely because a couple of people who had seen the draft strongly encouraged me to, despite the awkwardness. There's a broader conversation I hope people in EA will have around these issues, and perhaps this can be a helpful contribution. Because I don't feel best-placed to judge that, in the end I shared it with a few people, giving them permission to share on, or otherwise do whatever they thought was good with it. It seemed like there was a good chance that it would end up on the EA Forum, and I'm now informed that that's likely to happen (although at the time of writing I don't know who will post it, and I'll probably never know exactly whom it was shared with at this stage).

This post tries to capture my personal views and my own thinking rather than trying to convey something like a consensus view, or making claims about the correct view. A lot of the individual bits of thinking aren't original to me, but are based on ideas I got from comments or conversations with other people. So credit for any insights is diffuse; blame for flaws in the synthesis remains entirely mine.

Introduction

The FTX collapse has caused a number of updates for me. I'd thought the story was something like "group of smart folks motivated by ~EA principles throw themselves into a startup and become industry leaders". But now the story looks more like "group of smart folks motivated by ~EA principles throw themselves into a startup and become industry leaders **by committing fraud**". This is horrifying.

I've spent some time grappling with what I think the appropriate updates are, and wanted to give an outline of how my thinking has changed. (My thinking is based on private information as well as public information about FTX — in particular, I talked significantly to the Alameda founders in the early days, and to the Future Fund staff in 2022, and I was a Future Fund regrantor — but at this point I'm not sure the private information changes much.) In the next section of the post I talk about what I know and guess about what happened; in the following section I talk about general cross-cutting updates in light of this; and in subsequent sections I draw out some implications for different types of work. I'm here generally thinking about work motivated by longtermist EA considerations; I haven't thought about how much this will translate to other areas.

A lot of this post is tentative. I don't feel like I'm at reflective equilibrium. I'm likely to be missing important categories of updates, and be overblowing other parts. I'm sharing now because I think it's a helpful thing for us to be collectively considering; and because I hope that this could prompt useful thoughts in people reading (building on or disagreeing with what I'm saying).

Also: writing about updates is hard. These are all somewhat complicated topics with real tradeoffs. My views are normally based on lots of interwoven factors, and I don't get to fully convey those, although I've made a serious attempt here to gesture. So — these are all directional updates. I think many people should make updates in the same direction, but depending on the starting position you had, you might already have believed the main claim, or still not totally buy it after the new evidence.

After a bit of consideration, I've added in numbers trying to give a crude representation of my takes on different claims before and after my updates, on a 0–10 scale. These numbers were given quickly and shouldn't be taken as too meaningful (it's an undefined scale and I'm often vague about what the statement they're attaching to even is!). I expect them to be less useful to most readers than the information about the direction of updates and why; at the same time I thought it was good to try to be a bit more transparent about where my thinking is currently at.

I want to note that this isn't just academic. My views on what's right have shifted, and consequently I'll orient differently in my views on what I or others should do. It's possible that with further reflection and experience I'll move further in these directions — or decide that I'd overupdated, and move my position back. We'll find out!

Iteration is slow enough that I think it's correct to make updates large enough that they will be overcorrections a reasonable fraction of the time. There may be pragmatic reasons to make the corrections to actions bigger than the corrections to views (to experiment faster and hone in on the optimal policy sooner), or smaller (if oscillating policy would be disruptive); I'm not really trying to account for either of these in the policy updates I discuss, which are more like “my current guesses about optimal policy”. I should also remark that something that this *isn't* doing is deliberately leaning into a new perspective in order to feel out its strengths — I think that this can be a really helpful move, and is one that I used in the process of writing this post (and it's

possible that the vestiges of that are still affecting my judgements), but the strengths of views represent something more like my current all-things-considered best guess about where I'll end up.

To a great extent, any individual's views here don't matter that much for the community as a whole. I don't think people should care that much what I think right now; rather my hope for this post is that it will get people thinking about the range of possible updates they might want to be making, and in some small way feed into further and deeper processing of these questions.

What happened?

What do we know?

My understanding from the complaints from the US regulators is that four FTX/Alameda executives — Sam, Caroline, Gary, and Nishad — committed one of the biggest frauds in history, by advertising FTX as a safe place to trade and separate from Alameda, while simultaneously comingling funds and not subjecting Alameda to their risk limits. There are a number of other charges (that I don't fully understand the implications of). This came to light after FTX suffered a bank run while Alameda was deeply in the red. We don't know what the outcome of the bankruptcy proceedings will be or how much money will be recovered, but it looks like a lot of depositors lost a lot of money.

At least Sam, Caroline, and Nishad were significantly involved in the EA community. Sam had publicly expressed a desire to make money in order to give it away; Sam had also publicly talked about the importance of risk neutrality.

As discussed in [this article](#), a number of people in the EA community had concerns about Sam's business ethics, risk tolerance, or management. This information was not broadly disseminated. At the same time the FTX Foundation became one of the largest funders in longtermist EA, and Sam became one of the most prominent faces of EA (he was already one of the most prominent faces in crypto).

Some guesswork

I don't have reason to think anyone beyond those four knew about the central fraud, and my best guess is that nobody did. (I don't really think that detailed public speculation about who knew what makes sense when there are ongoing criminal proceedings; we're better waiting for the courts; nonetheless my best guess is that this isn't a case where lots of people knew, and that this is worth stating at least once.)

Two questions which are important for my updates are:

1. Why did these people commit fraud?
2. Why did the EA community relate to them the way it did?

I'm unlikely to ever know the answers to these for sure (especially 1), but they're central enough to the process of working out what if anything needs to change that I'm going to have to speculate.

On 1, some guesses:

- Probably they wanted to make a lot of money
- Probably this desire was at least partially driven by wanting to give a lot of money away
 - Support: They talked about this a lot; they spent time engaging with EA; other people in EA seem to have coded them as “on our side”
 - Support: They gave away quite a lot of money
 - Support: They might reasonably have expected the risk of one of the four of them whistleblowing to be lower if everyone understood that they were doing this with noble goals than for self-enrichment
 - Against: They didn't give away that much money, compared to what they spent on [sponsorship rights for esports](#) or [naming a sports stadium](#)
- Probably they were at least partially driven by wanting to personally become rich and powerful
 - Support: This is a very common motivation among humans
 - Support: Sam's behaviour with political giving seems like it was optimized in part to make himself influential (compared to e.g. if he'd just given money away to his foundation and not been personally involved in any decisions)
 - Support: They could have told themselves plausible stories about how becoming rich and powerful would be good for the world
 - Support: The strategy makes more sense as a gamble to personally become rich and powerful than as a gamble to do good in the world
- Probably they didn't set out to steal customer money in a Ponzi scheme
 - Support: this would have been particularly stupid (almost certain to end badly)
 - Support: this would have been a particularly blatant ethics violation (less likely to get them all to agree)
 - Support: this isn't necessary to explain their behaviour, compared to a story where they borrowed customer money (illegally and unethically) in order to trade with it, with the expectation of making money and giving it back with nobody the wiser
- Probably they underestimated the risks that they wouldn't be able to pay back customer money
 - Support: Sam didn't have a reputation as a careful risk-taker
 - Support: It's a common bias to not look closely at things one feels uncomfortable about
- Probably this wasn't just incompetence
 - Support: This is a wildly large amount of money to lose track of
 - Support: It sounds like deliberate choices were made to exempt Alameda from the risk limits

- Support: Some of Sam's public communications seem like there were straight misrepresentations compared to what it sounds like his private knowledge at the time was
- I have little idea:
 - How much this was a surprise to them as past incompetence caught up, versus a calculated risk having turned out poorly
 - *How* skewed their thinking was on the magnitude of the risks
 - How much they conceived of what they were doing as fraud (or when), and were able to look at the risks they were exposing their customers to, versus not letting themselves consider this
 - How much in their heart of hearts they believed they were doing the right thing
 - To the extent they believed they were doing the right thing, how much this was because they thought they'd empower others with donations vs thought they'd gain power themselves and make good use of it
 - How much other people at Alameda/FTX were complicit in what was happening

On 2, some thoughts/guesses:

- Even more than for 1, probably there's no single clean answer to this — we're talking about many actions across many people, who will have each have had many factors influencing their choices
- Probably conflicts of interest affected how some people thought about Sam/FTX
 - Support: priors
 - If someone approaches you and says "we'd like to give you money" or "we'd like to put you in charge of a large pot of money", it's typical for your brain to track that as a good outcome
 - If you're aware that if you came to believe they were unethical that would mean you shouldn't take their money, this provides an incentive not to look closely at the question of whether they were unethical (since by presumption in the status quo you will accept)
- Probably Sam's tribal affiliation with EA was a significant factor in people overlooking yellow flags
 - It's often easier to round things off into simple categories. The ability to put Sam & co in the category "EAs" may have made it easier to think "oh things that they do will be good" and not look carefully at the actions one could observe
- Probably some people who had little or no direct access were deferring
 - Attitudes something like "it seems like the legit EAs whom I know enough to have trust in think that he's good, so they've presumably checked this out enough to not be concerned" (although perhaps not so explicitly articulated)
- Probably social incentive gradients inhibited sharing negative opinions, even when people did notice issues
 - If someone thinks that they might at some future point want funding from a major funder, and that speaking critically might reduce their chances of getting that funding, they may rationally abstain from offering criticism, even in informal settings

- (This can be true even if the funder would never choose not to fund someone just because they were critical — because funding decisions often rely on judgement calls, making a credible commitment to this effect can be hard. & I didn't see any significant attempts from Sam/FTX)
 - We know of at least some cases where a reputation for vindictiveness on Sam's part made [people unwilling to share negative information without requesting it not be shared on](#)
- Probably a consequentialist desire to protect the reputation of EA made people reluctant to share negative takes after Sam was publicly associated with EA
 - If people can predict that the stories that would be written off the back of sharing negative anecdotes about Sam would be overall unfairly negative to EA, they may have wanted to stop those anecdotes reaching journalists (even if they think that the anecdotes are true and it would be good if more people knew about them)
- Perhaps a consequentialist desire to protect FTX's reputation made people reluctant to share negative anecdotes
 - If people think "look I've heard about some non-ideal behaviour, but nothing seems too bad, and if Sam's company does well he'll probably end up giving more money away to things I care about", that's an incentive to not propagate things that could lead to reputational harm and hence his company having less money to give away
 - (I'm not very confident about how prevalent any of these dynamics were, but I feel particularly unsure about this one)
- Perhaps the above dynamics inhibited any serious efforts to get anything to change
 - If negative information didn't propagate, it's reasonably likely that few if any people had all of the substantive negative evidence that was available somewhere
 - If people could perceive the social dynamics suppressing sharing of negative information, they may have felt powerless to change anything

General updates to my thinking about what's important

My updates are of two types, factual (i.e. the world is now different, so the best actions are different) and epistemic (i.e. processing this evidence I have different beliefs about the world than I used to, and in retrospect I wish I'd had those beliefs all along). [I'm not sure if there's a standard terminology for this distinction.]

First, the factual updates. These aren't necessarily the most important, but they are the clearest, because the situation the longtermist EA community finds itself in is different. Probably these are less debatable — factual updates shouldn't depend significantly on the models of the world you started with. Three major threads are:

- **Less money** — There is significantly less money available
 - "Money is tight": Agree 2/10 → 4/10

- **Brand** — EA/longtermism has a lot more media attention, and will have a serious stain on its reputation (regardless of how well deserved you think that is)
 - “How good is EA’s reputation?”: 8/10 → 5/10
- **Distrust** — My prediction is that if we polled the EA community, we’d find EAs have less trust in several institutions and individuals in this community than they did before November. I think this is epistemically correct: people should have less trust in several of the core institutions in the community (in integrity; in motives; in decision-making)
 - Why?
 - For many people the story (I presume) looks like “these people and orgs vouched for someone who turned out to be fraudulent”. This is compatible with “they were all duped”, but is also compatible with “something bad was going on here”; the reasonable response given lack of knowledge about which applies is to reduce trust!
 - NB I have had a lot of access to what’s happening in these groups, and this generally *hasn’t* decreased my trust (with the exception of the type of update I discuss in this post), but I’m conscious that nobody in leadership just gets to assert “these things are trustworthy” and have people know that they should believe them. Of course I also had somewhat nuanced takes on things I would and wouldn’t trust people on before the FTX collapse, so not updating too much doesn’t mean “I trust these groups in all things”.
 - [A salient question for people may be “(when) should we trust Owen?”; again I don’t get to just assert answers to that question. I may share my takes at some point to help people make sense of where I’ve been coming from, but that’s not something I’ll do here.]
 - My guess is this will change the way EA orients and thinks about accountability and structures in these organizations
 - “How much distrust is there in EA leadership and institutions?”: 3/10 → 6/10
 - (I’m afraid in not-related-to-FTX ways I’ve contributed to a recent move up to 7/10)

Second, the epistemic updates. (Readers might have had different models beforehand, and so update differently.) Several of these blur together, but salient ones include:

- **Exceptionalism** — Unwinding some updates I’d made in light of apparent FTX-success:
 - On EA exceptionalism (it now somewhat less appears that a competent group of EAs could expect to do well in arbitrary industries)
 - Relatedly, on the ease of making a lot of money (this means that my estimate of future funding streams goes down beyond what’s implied by the immediate cut in funding)
 - “How exceptional are EAs, compared to other social groups?”: 5/10 → 3/10
- Updates coming from the question “what might need to change in the EA culture to avoid someone deeply involved in EA causing a large catastrophe in the future?”. (Noting that I don’t know how exactly how well-grounded in current EA culture the folks in question were, but they were pretty involved in parts of it.)

- ***Dangerous ideas*** — We should be more worried that aspects of our memplex systematically increase the risk of people taking extreme actions that are harmful
 - Reasons for update:
 - Most intellectual communities don't have people affiliated with them that cause harm on this scale
 - While we might think the best version of our ideas won't lead to such harms, there are lots of ways to come apart from the the best version, and we should care about trying to make our culture robust to people only getting part of it
 - It's notable that compared to most of society, our most central ideas are potentially radicalizing:
 - [EA is about maximization, and maximization is perilous](#)
 - [We might live at a critical time in history](#)
 - (hence the implicit message "this stuff is much more important than anything normal in your life")
 - This mostly isn't directly decision-relevant, but makes me put more weight on other updates in this cluster (and could lead to further downstream conclusions)
 - "How big are the potential downsides of spreading our ideas?": 4/10 → 7/10
- ***By the book*** — The robustness that comes from doing things by the book seems more important
 - i.e. we should err further towards following typical best practices, especially in cases where the potential downsides are large
 - Reasons for the update:
 - It seems like FTX had a "move-fast-and-sort-out-process-later" culture.
 - While it's entirely possible to operate in that culture and not do anything wrong, and many people do so, I think that a culture which wants people to do things by the book will be less tolerant of groups which are actually violating important principles
 - People embedded in a culture which wants people to do things by the book may be less likely to think of plans which violate important principles in the first place
 - If EA had had more of a "do things by the book" culture I think they would have been more sketched out by FTX; also perhaps to the extent FTX executives were influenced by EA perhaps they wouldn't have ended up committing fraud in the first place (or perhaps they would just have associated with a different subculture)
 - This relates to ***Exceptionalism***; if you really regard EAs as exceptional, you may be less willing to challenge deviations from the status quo (because presumably they're improvements!)

- (I have also had personal reasons for updating towards doing things by the book over a similar period, and I'm not able to fully untangle this update)
 - "We should do things by the book": Agree 6/10 → 9/10
- **Uncompromising utilitarianism** — We should be more worried about people orienting to utilitarian arguments in absolutist ways that don't admit other heuristics
 - Reason for the update: While I don't know the details of people's motivations, it seems likely that a utilitarian desire to make money to give it away — and an intellectual commitment to the importance of risk neutrality — may have been important components driving the terrible decisions at FTX/Alameda. (In other words, naive utilitarianism is a real issue even for very smart people who are aware of all the normal arguments against naive utilitarianism.) Consequently:
 - I should be more worried when I get uncompromising utilitarian vibes from someone. When I drew [this diagram](#), I didn't really know how to compare the scale of the top-right and bottom-right quadrants; I now think it might be appropriate to put a stylized "-1000" in the bottom-right.
 - It's more important that our culture reflexively pushes back against naive utilitarian takes.
 - If someone has an argument X that A is good for consequentialist reasons, but it violates some commonsense heuristic Y, the burden of proof (at least at current margins) should be more on the side of explaining why the heuristic is inapplicable in this case rather than providing an argument that A is bad for consequentialist reasons. (I think in fact heuristics often have a consequentialist grounding, by encoding wisdom about diffuse indirect effects, even when these are hard to reason about directly.)
 - Although I had thought that everyone agreed on this already! We have lots of articles talking about why naive utilitarianism, or maximization, is bad. But I now think it's more important that this is an essential part of the culture; perhaps by socially penalizing arguments which are made just with a simple consequentialist model, without acknowledgement of what various heuristics make of the conclusions. (And not just left to individuals to read some articles and work out the implementation for themselves)
 - "How important is naive utilitarianism as an issue to address?": 5/10 → 9/10
- **Tribalism** — I'm more worried that people identifying as EAs is net destructive
 - Reason for update: I think that it creates bubbles where people feel trust because they're in the same tribe, without evaluating whether the things actually make sense. I think that that can be helpful at times, but is probably harmful overall if it leads to inappropriate trust.

- This is pure speculation, but: it seems quite plausible that tribal affiliation with EA led to the four people centrally involved in the fraud not whistleblowing (in this scenario I'm imagining they didn't frame it to themselves as fraud)
- I also think Sam's tribal affiliation with EA was likely a factor in people in the broader community overlooking yellow flags. (Because while plenty of non-EAs thought things were OK, including major investors, it's also the case that various people thought that his business practices looked sketchy; and I'd have expected more of such a reaction within EA absent Sam being coded as "one of us".)
- NB this is separate from the question of whether EA community building is net negative
 - Though I think that that's an important question and I'm not confident in the answer!
 - The prima facie plausibility is that we've been good at pointing people towards high-leverage areas and not so good at getting them to do a good job in those areas.
 - My best guess is that it hasn't been net negative, but I do think it's a real possibility, and that the FTX collapse is a data point which makes it more likely:
 - "EA community building has probably been net negative": Agree 2/10 → 4/10
 - "How much attention should people give to the question of when community building is net-negative?": 3/10 → 7/10
- Rather my idea is that more community building could be done in ways to encourage adoption of key EA ideas without adoption of an EA identity
 - Basically I think that EA-as-an-identity is a crutch which can help community-building (easier to get people bought into something when there's a clear thing to be brought into), but it actually makes the quality of the community building worse
 - This is partially because it encourages simplistic rounding-off and encourages people adopting a whole belief-cluster rather than considering each candidate view separately; and partially because one simple rounding off of EA is that it's about maximizing for doing good, and it's easy for that to lead to people engaging in a way where it tries to eat their whole lives
 - I think we should at least lean significantly away from it for a while and see what happens — outreach could focus on helping to give people tools for thinking through how to do good, with less focus on existing identities or communities

- One simplistic suggestion: taboo “EA”, “x-risk”, and a few other keywords from outreach talks; or at least relegate them to the appendices. I think this can’t really be the correct strategy but might be good as an experiment or jumping-off point for brainstorming
- “EA as an identity is bad”: Agree 3/10 → 6/10
- Another disjunctive approach to reducing the harms from tribalism would be not to reduce the degree to which people *identify with EA*, but to get them to change what that means to them
 - e.g. Encourage habits of making sure that you deeply understand the fundamentals of a project before you agree to join it, and don’t just take this on trust
 - (People could still work for projects just because they’re paid well, but they perhaps shouldn’t do something for impact reasons if it doesn’t fully make sense to them)
 - I’m not sure how fully this would work, because social and power dynamics would mean that people would be incentivised to trick themselves into believing they really understood something (when it was cool to do so), even if at some level they didn’t. Still, it might help.
- Updates coming from the question “how could the rest of the community more healthily have related to FTX?”, and related thoughts
 - **Conflicts** — I’ve moved towards thinking conflicts of interest, broadly understood, are frequent and really guide people’s thinking
 - Reasons for the update:
 - I think “conflicts of interest” is the simplest explanation for lack of dissemination of negative information about FTX
 - In some cases this may have manifested as people making deliberate choices not to disseminate negative impressions they had, for fear it would be bad for them personally, or bad for ~EA / the world
 - In other cases it might look like people *being resistant to the belief* about the bad thing in the first place, because part of them tracks that if it’s true, then they should disseminate it (and that would be bad for the above reasons)
 - I’ve found it hard to look straight at ways in which I or central parts of the EA community may have been messing up or causing harm. My brain keeps on wanting to slip away from it and go back to thinking about other things.
 - (This is pretty unsettling, when I let myself pause long enough on the phenomenon to make that observation)
 - There’s a clear conflicts-based account of what’s happening here: I’m in a position of power and influence

with respect to the community; it's convenient for me not to rock the boat

- I've found it weird how much disparity there appears to be in views about how much change/adjustment is needed
 - Some of this is probably because people just aren't talking much, or by differing amounts of information on the situation
 - I think at least some of it is likely explained by people being governed by conflicts of interest not just truth-seeking
 - For people who stand to lose in a shake-up of the status quo — or who might want resources from people who would stand to lose in a shake-up of the status quo (which is many people, since presumably the people who have lots of resources right now on average have more to lose than gain in any shake-up), it may be simplest to not give it that much attention
 - Vice-versa, for people who might stand to gain in a shake-up of the status quo, there could be an incentive to look for things that really need to change
 - “Conflicts of interest are a big deal that we should go out of our way to adopt and develop best practices for”: Agree 3/10 → 8/10
 - NB I'm not sure whether I'll think standard best practices are adequate. I more think: “our whole thing is being more right than other people about how to do good; this could be a major distortion in decision-making aimed at doing good; we should work out how to handle this well”
- **Integrity** — I think that upholding consistently high standards of integrity is particularly important
 - Reasons for the update:
 - I think that even if there hadn't been fraud, our values are the key part of what makes the EA community special, and it's a bad long-term move to hold someone up as an example without talking about their weaknesses
 - I think it would have been higher integrity to publicly criticize FTX while taking their money, than just to take money and stay silent (if one had critiques)
 - Talking to someone I respect (and think does important work) in the aftermath of the FTX collapse, they commented that they didn't identify as EA, and thought that actually compared to other

communities built around morality, EA seemed surprisingly low integrity, and that this was painful to them

- They weren't *confident* that it was low integrity, as it's possible some of the issues were explained by increased transparency compared to other communities; though there was at least enough actual issues observed in core EA to be painful to them
- They gave a parable about what I'll call integrity creep: "You act in a way that violates integrity. I learn that you do so. I could report you, but if I do then there will be consequences that I don't wish you to suffer. I stay quiet and become complicit. Now my integrity is impugned. It weighs on me, this complicity, and I mention this to a friend, seeking absolution. Now my friend could report you too. They do not. Now they are complicit too ..."
- I think this is the kind of person we really need; I think it's a big problem if our culture is pushing such people away
- "EA is a high integrity community": Agree 9/10 → 7/10
- **Taking responsibility** — Diffusion of responsibility for cross-cutting issues for the EA community can mean nobody works on them
 - Reasons for the update:
 - It seems like there was little work on "what should the EA community do about Sam?", despite enough dispersed negative signals that in the abstract it probably deserved serious attention from someone
 - I have an impression that many people would have thought something like "probably EA leadership has this handled"
 - (Anyone who didn't and tried to do something might have hit social obstructions as discussed variously above)
 - I think people in "EA leadership" would generally have thought "oh I don't know if I can self-appoint to own this issue" as well as "Man I have a lot of important things to do, I don't know if this should be a priority"
 - The people whose remit comes closest to naturally including this is the Community Health team at CEA
 - But they weren't in a structurally strong position — their head reported to the head of CEA, who reported to the EVF boards, which included Nick Beckstead, who reported (in his day job) to Sam
 - "We need better mechanisms for getting people to take responsibility for things": Agree 5/10 → 8/10
- **Complicity** — Tacit tolerance of bad behaviour is a serious issue
 - Reasons for the update:

- I think [this article](#) does a good job of articulating “complicity” issues
 - It seems like many people in EA were aware of negative signals about Sam, while many others were unaware (and probably trusted him because he appeared to be trusted by the EA community)
- I don’t think that awareness of negative signals was necessarily enough that people should have turned down money from FTX (& charities can have legal obligations to accept money), but at least some people did turn down money, and I think it’s bad that others didn’t have the information to make an informed decision
- More generally, Sam was treated as a core member of the community, and held up as an example to emulate
 - I don’t think that the people holding him up like this knew about fraud, but I think people had some sense of the person Sam was; given evidence now I think we should be pretty negative on his character flaws, and not talk solely about his virtues in a way that could give people a halo-effect impression
 - To some extent I think this is just an epistemic update: people were naive in a sense of “maybe that’s what business is like; that’s how you become successful; I don’t know enough to criticize”
 - But also to some extent I think people should have known enough about undesirable behaviour to think “that’s not necessary; that’s enough of a black mark that we should hold this guy more at arm’s length”
 - This update mostly doesn’t have immediate action guidance, but rather acts as an importance multiplier on the issues above — it makes me think “it’s important to get to the bottom of the dynamics at play”, and “it’s likely that some significant changes will be warranted”
 - “EA complicity in Sam’s minor misdeeds [i.e. excluding fraud or anything that nobody knew about] was a serious issue”: Agree 2/10 → 7/10

Implications for object-level work

A lot of this work will be research-based, but I think it also includes building institutions, coalitions, improving decision-making, etc. My current thinking on the direction of change implied by the above updates is:

- We should **be a bit more positive on people doing crucial work within established institutions**
 - What does this look like?

- Encouraging people who want to do centrally-important work to consider doing this within established institutions (such as universities or government agencies), even when that would add friction
 - Why?
 - ***Distrust*** — These organizations have their own internal checks and balances, and integration into mainstream society could help to earn trust
 - ***Less money*** — In some cases this will provide opportunities to leverage money from other mainstream sources that wouldn't otherwise be available for work we regard as crucial
 - ***Exceptionalism*** — We should be a little less bullish on our ability to outperform standard institutions, and a little less sceptical when they have processes that seem not to make sense
 - ***By the book*** — This is in many ways the standard way to proceed; we should give some weight to that
 - ***Uncompromising utilitarianism*** — These institutions often have cultures which would serve as some safeguard against naive utilitarianism
 - ***Tribalism*** — Integration into mainstream institutions is pulling in the opposite direction to establishing a separate EA tribe
 - “Mainstream institutions represent good homes for some crucial work”: Agree 4/10 → 7/10
- We should **have a somewhat higher bar for funding things**
 - Why?
 - ***Less money*** — (straightforwardly)
- We should **consider lower salaries**
 - Why?
 - ***Less money*** — If money is tighter, margins change; if salaries were optimal beforehand, they should probably come down slightly (if you thought they weren't optimal beforehand you might not think they should come down)
 - “Direct work salaries are at risk of going too high”: Agree 2/10 → 4/10
- We should **care a bit more that plans look robustly good**
 - What does this look like?
 - We should explicitly encourage Theories of Change to look at things with several different heuristics, and to take note of which heuristics the plans look good or bad on
 - We should treat it as a sign to proceed carefully or look for a different plan if something gets flagged as bad by some heuristics
 - We should be sceptical and treat it as a red flag when someone appears to be blithely ignoring the warning signs given by some heuristics (and perhaps a little sceptical if they appear to be acknowledging the costs but moving forwards anyway)
 - Why?
 - ***Uncompromising utilitarianism*** — This is trying to build guardrails against naive utilitarianism into our culture

- **Tribalism** — Looking for things to look good from multiple perspectives should help to reduce the extent to which people are only inhabiting a single “tribal EA” mindset
 - “How much should theories of change give multiple perspectives”: 3/10 → 7/10
- We should **be a bit more positive on [research distillation](#)**
 - What does this look like?
 - We should strive to be very clear and explicit with ourselves and others about what we believe, and what the evidentiary basis for that is. We should acknowledge places where the case isn’t watertight, rather than looking for the wordings that we think will be most persuasive in the short term
 - In aid of this, we should:
 - Encourage researchers to spend time finding the cleanest framings and articulations of points that may be important
 - Encourage soliciting of reviews by critical peers, looking not just for accuracy but clarity, and taking the time to address concerns and improve the framing
 - NB I don’t necessarily mean the academic peer review system; I think when it works well this system can do a great job of serving this purpose, but:
 - a) one can potentially get the same benefits in other ways;
 - b) not-infrequently academic peer review does a bad job of this; and there are a lot of articles in peer-reviewed journals which aren’t at an appropriately high quality bar
 - Write in ways to engage with the frames that are used by outside experts
 - To stress test our ideas by forcing them into other frames and seeing if they still make sense
 - To cross ontological gaps well enough to invite critique from outside experts
 - Why?
 - **Exceptionalism** — I have more a bit default scepticism than I used to towards claims that are commonly believed in EA but which don’t have clearly-robust foundations
 - **Distrust** — This could help to put core object-level claims on legibly stronger foundations (or reveal them to be mistaken!), which could help to shore up trust in direction-setting
 - **By the book** — This could help to improve robustness by forcing people to do the work of sorting out exactly what the foundations are (and probably uncover some gaps in our thinking along the way)

- ***Tribalism*** — This should reduce the extent to which people come to believe things because of implicit appeals to epistemic authority, which I think can contribute to an us-vs-them mentality
- ***Integrity*** — This could help us to hold our heads high when communicating about key issues; we would have stronger foundations for whatever we would then believe
- ***Taking responsibility*** — Research distillation strikes me as something which is often important and often regarded as to be left for someone else to pick up. I both think it would be good to get more of this to happen, *and* I think that there might be something culturally good about getting to a place where people coordinate (by self-appointment or otherwise) to make sure that this type of ball isn't dropped
- “Research distillation is a top priority in key areas”: Agree 6/10 → 8/10

Implications for community-building activities

EA, longtermist and otherwise, puts quite a lot of resources into community-building — i.e. work aimed at getting more people seriously engaged with and working on crucial issues (in some cases this is branded as “EA”; in other cases, not).

My current thoughts on the direction of updates here:

- **Content (reading lists, talks, etc.) should:**
 - **Be a bit more positive on content from outside EA**
 - What does this look like?
 - Reading lists including material that isn't in an EA voice
 - Events with speakers who have interesting things to share from beyond EA
 - Encouraging people to read different things, not all read the same things
 - Why?
 - ***Exceptionalism*** — we should engage more seriously with the value and expertise the broader world has
 - ***Tribalism*** — we should not implicitly send the message “only EAs are worth listening to”
 - “How important is it make sure to include non-EA content”: 3/10 → 6/10
 - **Be a bit more tools-driven, and a bit less answers-driven**
 - What does this look like?
 - Spending more time on principles and thinking tools
 - Spending less time on “these are the most important problems”
 - Why?
 - ***Distrust*** — we should give people the tools to audit things for themselves, not ask them to take it on trust
 - ***Exceptionalism*** — we should be more open to the possibility that we're getting the object-level judgements wrong

- ***Uncompromising utilitarianism*** — we want to foster a culture of questioning, which challenges simple answers, and this equips people to do so
 - “Tools should be emphasised more than answers”: Agree 6/10 → 8/10
 - **Have a bit more emphasis on the value of looking at things from several perspectives**
 - What does this look like?
 - Intro material could embody this virtue, applying different heuristics and considering what they all say before suggesting an all-things-considered judgement
 - In conversations or talks, we could expect more of this — and push back if it doesn’t seem to be happening
 - Why?
 - ***Uncompromising utilitarianism*** — this is a social technology for avoiding naive utilitarianism
 - ***Tribalism*** — habitually looking at things from multiple perspectives discourages tribalism
 - “How important is it to explicitly encourage looking at things with multiple heuristics / perspectives?”: 3/10 → 7/10
 - **Focus a bit more on social epistemology**
 - What does this look like?
 - Discussion of incentives, groupthink dynamics, best practices for managing conflicts of interest
 - (So that content on thinking is not just about individual rationality)
 - Why?
 - ***Integrity / Complicity*** — If there’s common knowledge of what bad and good behaviour looks like, it could be easier for people to suggest solutions, to socially punish bad behaviour, and to socially reward good behaviour
 - ***Conflicts*** — more people knowing what the issues are with conflicts of interest and how to mitigate those means we’re more likely to avoid issues
 - ***Tribalism*** — awareness of tribal dynamics could make it a bit easier for people to consciously avoid them
 - “We should include content on social epistemology in core syllabi”: Agree 3/10 → 7/10
- **The vibe of community-building activities should:**
 - **Lean a bit further away from encouraging people to identify as EA**
 - What does this look like?
 - Less branding activities as EA
 - Less branding of people or organizations as EA
 - Why?
 - ***Tribalism*** — (straightforward)

- **Complicity** — If we're not sure that we can hand-on-heart recommend trusting people and orgs in the ecosystem, we may want to be shy about making implicit recommendations to that effect via EA identity
 - "We should actively discourage an EA identity": Agree 1/10 → 5/10
- **Lean a bit further away from "we have the answers" and towards "we're giving you the questions"**
 - Why?
 - **Exceptionalism** — We don't want people to believe that we have answers that we don't
 - **Tribalism** — We don't want to encourage belief in an EA tribe
 - "How important is it to make clear that we have more questions than answers?": 5/10 → 8/10
- **Send somewhat fewer in-group signals**
 - What does this look like?
 - Trying to minimize jargon or references to things that people in-the-know will understand and will be opaque to those who aren't
 - Less putting people in situations where they're cut off from their normal support structures
 - Less all-vegan catering or other expectations that everyone behaves in a particular way
 - NB I'm not saying that event organizers can't choose to make something all vegan if they have ethical views which mean this is important to them — but that this update should make people regard this as a bit more of a negative for the signalling effects than they previously believed (you might still think it's positive, just less so; or might always have thought it was negative, and now moreso believe that)
 - Why?
 - **Tribalism** — these signals encourage people to identify with the group
 - "We should strive actively to avoid in-group signals": Agree 3/10 → 7/10
- **Focus on building a culture which is high-integrity**
 - What does this look like?
 - Socially rewarding people for calling out integrity violations — e.g. times when someone communicates or implies one thing to some group when they knew (or should have known) the truth was otherwise; or times when someone makes a decision that wasn't made for the best reasons
 - Creating common knowledge that minor integrity violations are common, so that having some doesn't mean the person in question is bad (but that we should strive to reduce the number we have)

- I think this is important! Without this, it's a strongly attacking move to call out integrity violations, which means that people will refrain from doing it, which means that they're taking hits to their own integrity
- Why?
 - ***Integrity, Complicity*** — (straightforwardly)
- "It's important that people publicly call out integrity violations they observe": Agree 3/10 → 6/10
- **Focus on building a culture which treats consequentialist analysis as just one tool in the toolkit**
 - What does this look like?
 - Encouraging and rewarding pushback against consequentialist analysis of the form "X looks bad on Y heuristic"
 - Encouraging curiosity about what's going on when different heuristics or analysis conflict
 - Why?
 - ***Uncompromising utilitarianism*** — (straightforwardly)
- **Focus on building a culture which asks people to make sure they know who has responsibility for things**
 - What does this look like?
 - Encouraging attitudes of "Be very explicit what you're taking responsibility for and what you aren't"
 - Encouraging people to demand these attitudes of other people and organizations
 - Holding people and orgs to account for things that they're claiming responsibility for
 - Trying to restructure or replace or outcompete them if they're not doing an adequate job
 - Stepping in and trying to appoint someone (self or otherwise) to issues where there's no clear owner taking responsibility
 - NB I haven't worked out the boundaries of when this is appropriate and when it goes wrong, but I think at the margin our culture could do with a **lot** more of it
 - Being clear about which beliefs we're taking responsibility for trying to keep accurate by first-order analysis, vs where we're outsourcing our beliefs to other people
 - Why?
 - ***Taking responsibility*** — I think this is an important dimension to improve on, and I think grassroots culture which supports it would be a good step towards that
 - NB I'm not confident, but I worry that EA culture may be actively harmful on this dimension — encouraging people to think "oh, someone probably has this covered"

- “How important is it to be clear+explicit what one is/isn’t taking responsibility for?” 5/10 → 9/10
- **Structurally, community-building activities should:**
 - **Put somewhat lower estimates on the monetary value of outcomes or programs**
 - Why?
 - **Less money** — money and people are complements, so less money reduces the value of marginal community building
 - **Exceptionalism** — we should be a little more sceptical that people can achieve exceptional things
 - **Tribalism** — I think a culture which puts high monetary value on attracting people to that culture feels like it could inherently slide towards tribalism; the choice of valuation can actually affect the value of community-building, and very high valuations can drive the actual value down
 - [My personal estimates will probably come down by about a factor of 5]
 - **Be more transparent about these valuations and other tools for decision-making about community building**
 - Why?
 - **Distrust** — opening up the reasoning for external audit can increase trust
 - “How important is it to take the time to produce public versions of key analyses”: 3/10 → 7/10
 - **Scale down activities a little (or slow the growth trajectory)**
 - Why?
 - **Less money** — marginal activities or experiments may no longer clear the bar
 - **Tribalism** — the disvalue of tribalism means that it looks easier to cause harm with community-building activities
 - **Brand** — It could be good to consolidate some on the highest-quality activities
 - “There is too much community-building”: Agree 3/10 → 6/10
 - **Scale down salaries a bit**
 - Why?
 - **Distrust** — avoid the appearance of grift (especially important for meta work which are less legibly impactful)
 - **Less money** — If money is tighter, margins change; if salaries were optimal beforehand, they should probably come down slightly
 - “EA meta salaries are at risk of going too high”: Agree 3/10 → 6/10

Implications for central community coordination

There are some mechanisms — significantly events — for facilitating coordination among the centre of the community. This is a small part of total EA effort, but: (1) it's salient to me, since I've been to most of the "EA leaders forum" and "EA coordination forum" events over the last few years, have advised on their format and structure, was an organizer for the one last year, and have been involved in organizing a number of other small community coordination events; and (2) this is one of the domains which should plausibly be making big changes in light of the FTX Saga

My current thinking on this is:

- We should **lean a bit further towards professionalism**
 - What does this look like?
 - Event more structured around the things we want to achieve, and erring towards using more normal workshop/conference structures (like prepped sessions with goals) to achieve this, rather than relying on unstructured participant-driven content
 - NB I don't mean that I want to see the trappings of professionalism, or virtue-signalling over looking professional; rather I'd like people to spend longer meditating on what good things professionalism as normally practiced is achieving, and then making sure that we're getting those. I expect this to look like "leaning a bit more professional"
 - Why?
 - **Distrust** — making the structures more explicit and direct could help increase trust in the purposes and decision-making behind the events
 - **Exceptionalism / By the book** — I have more expectation that some good things would happen from using standard structures, even if I'm not 100% what they would be
 - **Tribalism** — if events project "we're all friends here" (as some have deliberately chosen to do), this puts some attention on the group of people rather than the topics, and could lead to inappropriate trust
 - **Brand** — the world is looking at us more, and since it's cheap for us to lean professional we should take the opportunity to signal that we're serious // this isn't just a social scene
 - "Community coordination events should primarily feel professional": Agree 1/10 → 7/10
 - (I *don't* mean "the same professional one finds all over the place; I mean "a deep true version of professional that we build and that is recognizably professional once built")
- We should **lean a bit further towards transparency**
 - What does this look like?
 - Going out of our way to let the EA community know about events that we predict they might be interested in the existence of, even when those events are invite-only

- Sharing information about the format, decision-making, impact-assessment, and output (if any) of the events
 - At minimum, doing this in some fraction of cases — the costs of doing this scale approximately linearly with the number of events it's done for, whereas the benefits are diminishing (sharing a few helps to make the general style of thinking transparent)
 - Why?
 - **Distrust** — It's more important to take moves to help justified trust. Transparency lets people know what's going on, and lets people audit things for themselves more (and so work out whether this is actually something they want to trust).
 - **Exceptionalism** — Transparency increases the likelihood of people pointing out improvements we could make (or taking things into their own hands and doing better things).
 - **Uncompromising utilitarianism** — Transparency is the type of thing where the benefits are pretty diffuse, but there's a commonsense heuristic in favour of it. I think the benefits of transparency get underweighted by explicit utilitarian takes, and I want to weight that a bit more highly.
 - **Complicity** — I think a general instinct towards transparency will make it harder to end up complicit in things we don't endorse, because we'll more reflectively be checking "would I be happy with this being transparent?"
 - "How much of a priority is transparency about community coordination mechanisms?": 2/10 → 6/10
- We should **consider creating mechanisms for anonymously sharing updates/impressions**
 - What does this look like?
 - Creating mechanisms to allow for members of a vetted group to share thoughts anonymously (either with the rest of the group, or publicly)
 - e.g. could do this for groups like "EAG attendees", "EACF attendees", "everyone on the EA Forum with >100 karma"
 - Why?
 - **Conflicts** — giving people a mechanism to express things without worries about what it will do to their social standing
 - **Complicity** — giving people who want to vent about a particular issue a way to do so in a way that credibly makes them not just "random aggrieved person on the internet"
 - **Distrust** — particularly if these were public, outsiders could audit them and reach their own judgements about what had been said (and hadn't been said)
 - (NB there might be legal reasons not to establish a platform which people may use for libellous statements; IDK)
 - "How much use should we make of mechanisms for anonymous sharing of impressions?": 3/10 → 5/10

- **Orgs should be very explicit about what they are and aren't taking responsibility for**
 - What does this look like?
 - *Ideally in public*, but *at least* via private coordination mechanisms, be very explicit about which outcomes or effects in the community or the world at large different groups are taking responsibility for
 - Inviting others to hold them to account for failure to do a good job of taking such responsibility
 - Not taking responsibility for things we don't have capacity for (without explicitly putting down other responsibilities)
 - NB there are some puzzles in exactly how this should be set up —
 - It's a big world; anyone can self-appoint to take responsibility for anything; sometimes it will seem that nobody who has taken responsibility is doing a good job and that someone else should step up
 - This is potentially a lot of information that needs to be kept track of. I'm not sure if we have the right information architectures to handle that information flow.
 - NB I don't mean for this to reinforce people protecting their turf — generally I've been wanting to see more competition in EA than we've had, and I'd still like that, even within areas that people are trying to explicitly take responsibility for
 - Why?
 - ***Taking responsibility*** — It should be transparent to people which things we are and aren't taking responsibility for, so that they can react accordingly
 - ***Integrity*** — I think a reason I'm particularly keen on this is that it's a kind of meta-level integrity: being very clear with oneself and others about what should and shouldn't be expected
 - “How explicit should orgs be about what they're taking responsibility for?” 4/10 → 9/10
- **Coordination mechanisms should facilitate making sure someone is taking responsibility for important things**
 - What does this look like?
 - Keeping curated lists of important things that nobody is taking responsibility for
 - A key function of community coordination events could be *trying to reduce things falling through the cracks* — finding important things that nobody is taking responsibility for, and seeing if someone good has capacity to take it on, else advertising the fact that it's a gap
 - Community coordination events could also be used as opportunities to *hold people to account* for things they previously said they'd take responsibility for
 - Why?

- **Taking responsibility** — It should be transparent to people which things we are and aren't taking responsibility for, so that they can react accordingly
 - **Complicity** — It's better not to allow ourselves room to have something important quietly get dropped because nobody dealt with it
 - "How much of the value of community coordination mechanisms should come from people to take responsibility for things?" 2/10 → 5/10
- We should **ensure that people can access some core discussions by application, not just by networking**
 - What does this look like?
 - Making sure that there are pathways such that if someone great but disconnected with current networks wants to be plugged in, there is a way for them to do so which looks like "demonstrate excellence"
 - This probably means some structures with a real quality filter which people can apply to (without committing to doing a certain type of work at a particular org)
 - This doesn't mean that every event needs to take applications; just that there need to be routes which go that way
 - Why?
 - **By the book** — This tickles my "seems an obvious normal sensible thing to do" sense
 - **Tribalism** — If the way people get access to exclusive spaces is (and hence is perceived to be) by networking, it creates an incentive for people to network; to focus on fitting in; perhaps to adopt the community identity
 - "How much do we need robust application-based mechanisms to access core discussions?": 5/10 → 8/10
- We should **lean a bit more towards legible invite criteria, especially for flagship events like Coordination Forum**
 - What does this look like?
 - Having clear criteria for invites such that who comes is less judgement-driven
 - Why?
 - **Conflicts** — For people determining invites, makes it less susceptible to conflicts; for people who hope to be invited, reduces the incentive-to-network
 - **Tribalism** — I worry that private events with opaque invite processes which feel like they're making claims to be where important things are happening could feed into a "us against the world" kind of mentality. This is a kind of tribalism where the tribe is narrower than "EA", but more like "insiders". Professional-feeling processes for invites could help to mitigate this.
 - "How important is it to have legible invite criteria?": 2/10 → 4/10
- We should **lean a bit further towards frugality**
 - What does this look like?

- Paying more attention to controlling budgetary costs for events (possibly at the expense of mild inconvenience for event attendees)
- Perhaps asking event attendees to pay costs (as a sanity check on the value of the event); I'm not sure when this will be a good idea, but it feels worth considering
- Why?
 - **Less money** — The optimal tradeoffs have changed, and we should be a little less in the mode of “look for ways to get good things to happen” and a bit more in the mode of “attend carefully to the associated costs”
 - **Brand** — Spending money on things that are luxuries is never a good idea, but if we're worried about people questioning motives for things, it's particularly helpful to make things legibly non-luxurious
- “How important is it to be frugal for community coordination events?": 2/10 → 5/10

How this affected plans for the Summit on Existential Security

For a few months I was collaborating with the CEA Events team to plan the Summit on Existential Security [LINK], which is in the broad category referred to here (more recently I stepped back from that). Since plans were already significantly in motion before the FTX collapse, some of the updates weren't very actionable this time around (e.g. the venue had been paid for, so there's relatively little to do from the frugality update; and we'd largely settled on an invite process).

Some updates which had some impact in this case:

- We've skewed the event format towards professionalism (within the envelope that made sense to us given what was already in motion)
- We decided to invest time in communicating more with the community (leading to that FYI post, and also causally connected with me writing this post)

Implications for governance

In many ways FTX looks like it was acutely a failure of governance. This has changed my views on the appropriate tradeoffs:

- We should **increase oversight of projects and decisions**
 - What does this look like?
 - We should make sure that substantive decisions are made with oversight — people who are checked in enough that if there were errors of judgement those can be caught (at least after the fact)
 - This probably means that most people should be reporting to either a manager or a board who collectively spend at least 100 hours per year on guidance and oversight (~5% of the time of the person doing the work)

- This in turn means that we need to make room for more oversight hours, and these will come out of the time budgets of people with high opportunity cost
 - There's an important trap to avoid here about oversight-for-the-oversight: providing high-quality oversight can be important but also tricky; it's hard to assess from the outside when oversight is working well, *except* when something goes badly wrong (when one thinks "oversight should have caught this")
 - If one judges oversight primarily by being negative when there's a big problem, this incentivizes people providing oversight to be overly conservative (since they don't get credit for big successes)
 - Some mechanism design might help here (e.g. if we had impact markets giving the people providing oversight some impact equity could help align their incentives towards the kind of judicious ambition we should want to see)
- Why?
 - **By the book** — This is the standard best practice way of ensuring that things are not going off the rails
 - **Tribalism / Exceptionalism** — We shouldn't assume that EAs are special and don't need this; moreover a culture which suggests that we have more important things to be doing feeds into tribalism
 - **Conflicts** — I expect that people running projects will often have their judgement significantly affected by things other than "what is best", unless there is some force holding them to account for their actions
 - NB this might not be transparent to them! I think the experience of being conflicted is often not even noticing that one is conflicted
 - **Distrust** — People would like to have justified trust that EA orgs are being run well, and good oversight is an important part of that
- "It's worth aiming for ~everyone doing important work to have 50+ hours/year oversight": Agree 5/10 → 8/10
- We should **increase transparency of governance**
 - What does this look like?
 - Publish details of what the internal governance/oversight processes at key orgs look like
 - Consider publishing details of important discussions
 - Make processes for determining who holds governance roles more legible
 - Why?
 - **Tribalism** — If people can see robust governance mechanisms in place, it weakens the bid to trust the culture and people involved
 - **Conflicts** — If people governing things expect to be held to account by the public, they'll do a better more consistent job of managing conflicts
 - **Complicity** — If people governing things expect to be held accountable, they're less likely to let other things slide when they shouldn't
 - **Distrust** — (straightforwardly)

- “How important is governance transparency?": 4/10 → 8/10
- We should **err towards doing more impact analyses**
 - What does this look like?
 - More frequently doing explicit impact analysis; investing in making it represent a fair all-things-considered judgement; and publishing it
 - (As opposed to individuals doing impact analysis in ad hoc ways which help them to have a coherent picture but are hard to convey to other people)
 - NB I don't mean “there should be an impact analysis for every project”; I think that these can easily get performative, and for the time investment that doing them well takes I think the information value won't always be worth it
 - Why?
 - **By the book / Exceptionalism** — Probably people are making some errors of judgement, and forcing things to be explicit could help to catch more of them
 - **Distrust / Tribalism** — This would help move away from implicit bids to trust decisions because of trust in the decision-makers
 - “We should keep explicit impact analyses frequent”: Agree 4/10 → 7/10
- Projects and orgs should **invite accountability primarily for whether they took responsibility for the right things, and how those things went**
 - What does this look like?
 - Orgs should be clear and explicit to stakeholders (boards; funders; perhaps the community or public) about what they're taking responsibility for
 - They should be assessed primarily on:
 - (i) Whether those are the right things to have taken responsibility for
 - (ii) How those things went
 - They should avoid motte-and-baileys of responsibility where when it's convenient they implicitly claim responsibility for large swathes of things, but if pushed retreat to “oh no really we're only doing this small thing”
 - Why?
 - **Taking responsibility** — If we want this to be a core part of culture (as I think we should), we need it to be supported by our governance mechanisms
 - “Being explicit about what people are and aren't taking responsibility for is a central issue in governance”: Agree 3/10 → 9/10
- We should **give less weight to straightforward consequentialist PR arguments**
 - (but more to “how would this look if everything was transparent?” and “what do normal heuristics say about what to do here?”)
 - Why?
 - **Uncompromising utilitarianism** — (straightforwardly)

- **Complicity** — I think that “sharing this information would have bad first order effects” is one of the stronger tools for suppressing information that often shouldn’t be suppressed, and it would be good to weaken this tool
 - **Integrity** — This approach should make it easier for people to interact with things without taking any hits to their integrity
 - **Brand** — I think that this could help the long term brand, as people do more auditing of everything that’s going on
- “How much should we weight straightforward consequentialist PR arguments?”
5/10 → 2/10
- We should **spread governance work over more people**
 - What does this look like?
 - Bringing more people onto boards or other governance structures, including:
 - People with outside governance expertise
 - People who should *learn* good governance (if joining an otherwise strong board)
 - Make choices about who to involve in governance more legible as far as possible, and less purely network-based
 - Why?
 - Straightforwardly more work to go around if we’re increasing the amount of oversight we invest in
 - **By the book** — The world at large has a bunch of wisdom about how to make sure things are well-governed, and we should be more proactive in making sure we benefit from this
 - **Distrust** — There’s no reason to keep power centralized in a few hands when there are many more capable people it could be shared between
 - **Tribalism** — Spreading this work out could help to make it clear that it’s just an important kind of work, not membership of some special club
 - “We should involve many more people in governance”: Agree 5/10 → 9/10