Incentives and helping people to speak up

Background

- Resources
  - Twitter thread describing situation:
    - https://twitter.com/geoffanders/status/1456379139988996105?s=20
  - Habryka bounty (LessWrong comment):
    - https://www.lesswrong.com/posts/XPwEptSSFRCnfHqFk/zoe-curzi-s-experience-with-leverage-research?commentId=4s5dYnFiBJ7FGXJzP
  - Leverage Research initial inquiry email:
    - https://www.lesswrong.com/posts/XPwEptSSFRCnfHqFk/zoe-curzi-s-experience-with-leverage-research?commentId=95AjAiib5KGKeJYgG
- Bad optics
  - Obviously looks like I don't want info to come out.
  - But, need to defend people who are afraid of Habryka and the rationalists.
- Challenge:
  - Understand what circumstance would cause as many true stories to come out, without unduly pressuring people and without causing acute distress or lasting harm where it can be avoided.

Building context

Goals

- [Understand tons of facts about the circumstance → understand the causal structure →] understand what happened → understand where there were harms → understand how to have those harms not happen again
- Have people be able to talk about their negative experiences:
  - Spiracularity
  -

Desirable states

- Can each side credibly signal non-retaliation?

Key concepts

- Credible signs -- Is there something you can do that credibly signals to the other side that X? Is there something they can do that credibly signals to you that Y?
  - **Could Habryka create a duplicate bounty for harms from the Rationality community?**

        - And could he explain how he would protect the people who would come forward from being attacked by the Rationality community?
- Trusted intermediaries -- Can you find people who both sides trust who can intermediate?

Perspectives / [person types]

- Person who had negative experiences, feels unsafe speaking out about their experiences because of expected retaliation from me / Leverage.
- Person who had [negative / mixed / positive] experiences, feels unsafe speaking out about their experience because of expected retaliation from Oliver Habryka or the rationalist community or EAs.
- Person who feels like conditional payment will cause them to unconsciously tune their account to what they think the bounty-offerer wants.
    - ...and they believe the bounty-offerer is a long-time attacker.
    - ...and they believe the bounty-offerer is angry crypto-rich people on LessWrong.
- Person who does not want to interact with people they perceive to have harmed them.
    - → ex-Leverager believes that they have been harmed by Leverage/Geoff
    - → ex-Leverager believes that they have been harmed by Habryka/rationality community
- Person who reports that financial incentives of any type interfere with their writing.
    -
-

Scenarios

-
-
- Do vs. do not want to be financially incentivized to speak out
    - Scenario A: Some ex-Leveragers were harmed, want to be financially incentivized to speak out.
    - Scenario B: Some ex-Leveragers were harmed, and do <u>not</u> want to be financially incentivized to speak out because they don't want to confront or interact with the person or group or entity that harmed them
-

Proposal ideas

- **Unconditional payment/reward:** Habryka guarantees that he will pay for any account from an ex-Leverage ecosystem member regardless of the valence.
    - Context: The current bounty (v2) has Habryka make a judgment call as to whether and how much to fund accounts/stories.
    - Effects:
        - (1) This strongly discourages people from writing to try to claim a reward if they do not trust Habryka.
        - (2) This creates a financial incentive for people to try to write more negative accounts than they otherwise would if they believe Habryka is more likely to reward those.
            - → this encourages some people to lie
            - → this encourages some people to lie to themselves so as to not have to lie but still be able to get the money
            - → [other things like this]
    - Problem:

    - Specific instantiation:
        - $5k is offered to any ex-Leverage ecosystem member for an account, regardless of valence or quality.
        - Problem: gameability by non-substantive account
            - Solution #1: neutral judge for substantitiveness, not valence
            - Solution #2: low enough financial incentive that it's not too financially costly
- **Intermediary judge:** Someone (e.g., Anna Salamon) judges [quality] of accounts, decides what to pay out.
    - Problem: affiliation with Rationality for those who do not want it
    - Possibilities:
        - Trusted individuals (e.g., Anna Salamon)
        - Trustable organizations (e.g., HR firms)
        - Coalitions (e.g., me + Habryka)
- **Gather more info about potential writers' preferences.**
    -
    -

~45 ex-Leverage ecosystem -- $5k each -- $225k
~20 nearby affected people -- $1k each -- $20k
maximum paid out = $245k for 65 accounts