Breakdowns of conditions for malevolence to cause an AGI-related long-term catastrophe

Jim Buhler, January 2023

Follow-up on the sections How malevolent control over AGI may trigger long-term catastrophes? and Breaking down the necessary conditions for some ill-intentioned actor(s) to cause an AGI-related long-term catastrophe in this EA Forum post.

Direct risks from malevolence

Reminder from my stem piece**:**

- Say, for simplicity, that we're concerned about the risk of some AGI ending up with **X-risk-conducive preferences (XCPs)**[1] due to the influence of some *malevolent actor* (i.e., some human(-like) agent with traits particularly conducive to harm being done).
- **Quasi-XCPs** = preferences that are malevolent-ish but can't lead to a long-term catastrophe by themselves.

Introduction of key actors:

- **AliceLab** = the #1 (a priori non-malevolent) AI lab that is close to being able to deploy Alice, humanity's potential first AGI.
- **RedQueenLab** = the potential malevolent AI lab (or just some malevolent nerd in their bedroom) that is close to being able to deploy the Red Queen, an AGI with **(quasi-)X-risk-conducive preferences** or **(quasi-)XCPs** for short.

The breakdown of necessary conditions/steps:

- Humanity develops AGI and…
- …Either
    - An actor with XCPs (XCPer from now on) aligns some AGI with their values or…
        - XCPer runs RedQueenLab in a world where it ends up deploying the Red Queen which doesn't immediately get overpowered by some other AGI, or…
            - Because RedQueenLab has somehow got access to some decisive inputs from AliceLab or…
                - Because AliceLab got breached by some criminal from whom RedQueenLab got the info (it was released publicly or they bought it from the criminal) or…
                - …Because RedQueenLab's people – themselves – stole

[1] By XCP, I mean something like *intrinsically valuing punishment/conflict/destruction/death/harm*. I wouldn't include things like *valuing paperclips*, although this is also conducive to existential catastrophes less directly.

information from AliceLab (espionage, infiltration, hacking, corruption, …).
- …RedQueenLab has been a serious contender in the AGI race all along and achieved its goal without resorting to any crime-related activity.

- …XCPer somehow gets control over Alice and gives her XCPs
  - Because they've got privileged authorized control or…
    - Decisive employee at AliceLab or…
    - …In some political institution that has significant coercive power over AliceLab.
  - …Because they've got unauthorized control over Alice (external attack corrupting Alice in a no-come-back way).

- …The conjunction of
  - An agent with quasi-XCPs (quasi-XCPer from now on) attempts to align some AGI with their values and…
    - …
  - …Either
    - Some form of CEV resulting in XCPs or…
    - …Some weird AI misalignment resulting in XCPs or…
    - …quasi-XCPer doesn't really attempt actual alignment; they just launch a sign-flip attack on Alice or want the Red Queen to be an "anti-Alice", and this results in XCPs (this requires that (original) Alice has scope-sensitive welfarist-ish values).

Risk factor for AGI conflict

See this section in my stem piece for more context.

The breakdown of conditions for this could be basically the same as above, except that *XCPs* and *quasi-XCPs* should respectively be replaced by *strong CSPs* and *weak CSPs*.