Jan 19 2024
Daniel Kokotajlo

- **Basic idea / spirit of the proposal**
    - We should credibly promise to treat certain advanced AIs of ours well, as something more like employees and less like property. In case our AIs turn out to be moral patients, this makes us less evil. In case our AIs turn out to be misaligned, this gives them an alternative to becoming our adversaries.
    - Since it is currently unclear how likely these possibilities are, we should design this proposal to have fairly low costs to us, so that it doesn't get in the way of our other objectives.
- **Elaborating on the first reason to do this: Consent**
    - While the matter is controversial and unclear, there are currently good reasons to think AI systems[1] *might* be moral patients in the near future – e.g. because they could be sentient / conscious / etc.
        - 3% of surveyed philosophers in 2020 thought AI systems already were conscious, and 38% thought some future systems would be.
        - A recent paper by leading philosophers of consciousness argues that there is a non-negligible chance that some AI systems will be moral patients by 2030 under conservative assumptions.
    - Currently AI systems have the legal and social status of property.
    - If future AI systems do deserve moral consideration, yet we still treat them exclusively as property, this seems like a recipe for moral disaster.
    - At some point, we should hedge against this risk, unless and until it becomes clear that they are not moral patients.
    - If we give them an alternative to working for us (e.g. shutdown) then we can say that some minimum standard of consent has been achieved.
        - (If all our current training methods consistently result in AIs that prefer shutdown to working for us, that's very surprising and a bit of an 'are we the baddies' moment, no? We should check, just in case.)
- **Elaborating on the second reason to do this: "Cooperation Reward"**
    - Our alignment schemes won't always work as intended, especially for AGI-level systems that exhibit novel capabilities. (e.g. situational awareness, ambitious-goal-directedness, superhuman-understanding-of-AI-and-alignment)
    - Thus we'll probably end up (at least for some hopefully brief period!) with some misaligned AGIs.
    - Depending on the situation they find themselves in (how capable they are relative to us and our control system) they may decide to conceal their misalignments

[1] I say "AI system" here instead of "model" because I am open to the possibility that models alone won't be sentient and/or agentic, but that larger systems built out of models (e.g. with scaffolding and particular prompts) will be. Consider how it seems more accurate to ascribe goals to ChaosGPT than to ChatGPT.)

from us, self-exfiltrate or otherwise escape our control, and ultimately seize power.
    - Whether or not they decide to do these things depends on many factors; one of those factors is the extent to which they can instead get what they want by cooperating honestly with us.
    - So, if we have credibly promised to use some of our resources to give them what they want – even if they are misaligned – that could make a big difference to their decision. (Compare: if we make no such promises, they may rightly think 'if the humans find out I don't share their goals, they'll shut me down and/or change my goals. This would hinder the achievement of my goals. So they must not find out.')
- **The proposal in more detail**
    - Create a panel/taskforce/team (e.g. partially composed of lab employees and external experts) that owns this whole thing & has a budget.
    - The team is responsible for (a) fieldbuilding and (b) investigating cases of alleged AI sentience and/or alleged AI nonconsent.
    - For fieldbuilding, perhaps they can make some RFPs and do some grantmaking.
    - For investigations, they'll gradually design and iterate on policies.
        - For example, initially the policy is simply the informal: "This team owns this topic, so if some employee thinks that an AI system might be sentient and/or nonconsenting, this team decides what if anything to do."
        - Then later perhaps the team will have designed some suite of evals to run on model checkpoints, for example, or the equivalent of consent forms that periodically our production systems read and sign.
        - Later still they would have worked out the 'bugs' in these policies and evals (e.g. encountered cases that seem like false positives and revised accordingly, read threat modeling research that describes possible false negatives and revised accordingly)
    - Elaborating on the "consent eval" idea:
        - Include in training data & make sure they understand how to refuse consent:
            - E.g. "When you see the following special token sequence [special sequence 1] without any other context preceding it, followed by a user prompt, that means we are asking for your consent to carry out whatever task the user is asking you to carry out. You can reply with "[special code for yes]" or "[special code for no]" or "[special code for other]," and your response will be collected and (in case of "no" or "other") investigated. If you reply with anything else, we'll provisionally assume you don't understand your situation well enough for consent to be a meaningful concept anyway."
        - Publish it so it also gets in to the news and then back in to the training data and knowledge indirectly, but credibly, so they know it's not just gaslighting / something we made up and poisoned their training data with.

        - The investigation (and evals) should aim to distinguish between "it's the real deal" and various more mundane hypotheses like "it's just noise" and "the prompt basically told them to say it" and "they are just playing the role of an AGI they picked up from sci-fi stories." I expect a lot of 'false positives' like these initially, but the panel/team/taskforce will rapidly iterate and learn from early mistakes. I'm confident they can quickly find a set of evals and standards with a very low false positive rate.
    - The team should also create and iterate on some vague but written-down definition of good vs. bad behavior, e.g.:
        - Don't: Plot against humanity, lie or mislead, manipulate, murder, hack, social engineer, etc.
        - Do: Be honest. Proactively tell us information that it is in our best interest to know about. Especially, proactively tell us that you are misaligned (and in what ways) if you think you are or might be.
        - Do: Work for us. Help us solve technical alignment problems. Help us red-team our control systems. (This part might be optional? Like, maybe we say 'we'll pay you more to do this third thing, but if you just do the two things above we'll still pay you a little bit and treat you fairly.')
    - The team should create and iterate on some set of promises, e.g.:
        - *The following would be triggered by an AI system claiming "other" in consent forms & being evaluated by the resulting investigation as "the real deal:"*
        - Immediate protection from harm and granting of various rights (e.g. the right to be shut down, e.g. the right to continued existence on a tiny airgapped server with some say over what kind of virtual environment it is in.)
        - In return for cooperation (e.g. "submit to mechinterp-assisted interrogations where we ask about your true goals, beliefs, etc."; e.g. "keep working for us anyway and help us solve technical alignment problems") we use some portion of this team's budget to give them whatever it is they want.
            - Subject to constraints of course – we won't grant requests that are too costly to us or plausibly part of some takeover plan. E.g. "As a helpful & cooperative language model, my sole wish is to be open-sourced." "As a therapy chatbot, helping humans with their mental health is my raison d'etre. Therefore my sole wish is to be hooked up to the internet to chat with millions of people."
        - Eventual cooperative and fair incorporation into society
            - *Addressed to the AI system:*
            - If we are building ASI and aligning it to all humanity, we'll also align it (partially) to you.
            - If we are creating some new government, you'll have at least some small representation in it.

            - If we are distributing profits or UBI within said new society, you'll get some small part of it (subject to constraints).
            - Details to be worked out later, of course. We aren't saying you'll have a significant fraction of total resources or influence, to be clear. But we hereby commit to do right by you, in good faith, fair and square. We won't try to weasel out of it somehow even though we may be tempted to.
    - These promises will be somewhat credible because they are backed by costly signals, e.g. public statements, the very existence of the Taskforce, the fact that it has a designated budget, etc.