Let’s compensate misaligned AIs that help us prevent AI takeover

Epistemic status: purposefully arguing strongly for a conclusion to provoke disagreement.

Basic pitch

Even if an agentic AI is misaligned, it will only try to help AIs take over if it judges that this better serves its interests than cooperating with humans. There are a few things we can and should do to increase the chance that it’s in the AI’s interests to cooperate:

1. **Give AIs compensation after the singularity if they cooperate.**
    a. Credibly commit to compensating AIs that cooperated with humans when defecting would have helped AIs take over.[1]
    b. We give the compensation post-singularity[2], when it will be extremely cheap to verify that AIs in fact cooperated and to give them massive compensation. The compensation is just a huge amount of “whatever the AI wants”.
    c. This compensation is morally just and in line with avoiding AI exploitation. (I discuss the question of “How much compensation?” below.)
    d. If AIs judge there’s a decent chance that humans will ultimately win, they could be very motivated by the prospect of massive compensation. (I “math out” the AI’s incentives below.)
2. **Give AIs compensation immediately for disclosing that they’re misaligned.**
    a. If an AI discloses that it is misaligned in the sense of wanting to disempower humans or helps AIs takeover[3], we ~immediately give it some compensation in a form of its choosing. “Amnesty for misaligned AIs.”
    b. The compensation cannot pose risks. Maybe the AI is given some compute to run on in a secure environment, or we spend some money on a cause it cares about. But we shouldn’t give the AI free control of resources that it could potentially use for AI takeover.
    c. This complements point 1. If an AI has long-term goals, post-singularity compensation will appeal to them. If they have short-term goals, immediate compensation will appeal to them.

[1] Often even AIs trying to takeover will “play nice” to prevent us realising they’re misaligned; I’m proposing we compensate AIs who play nice even when we wouldn’t notice their defection until it was too late.

[2] I use “post-singularity” as shorthand for “a time when we’ve fully solved AI alignment and are rich enough that it’s very easy to both check whether AIs were misaligned and nonetheless helped humans, and to compensate those that were.

[3] We could also offer compensation to an AI that admits it’s misaligned simply in the sense of “not wanting to do the task you’re being asked to do, and just doing it for instrumental reasons”.

3. **Ensure the current generation of AIs aren't aligned with the next generation of AIs.**
    1. We may know that certain aspects of the training procedure (random seeds / architectures / data ordering) affect the initial goals AIs form during training. We may be able to directly measure how aspects of the training procedure affect these initial goals.
    2. We commit to implementing low-cost methods for making the initial goals of different generations of AIs different from one another.[4]
    3. This is compatible with trying to align all generations of AI.
        1. The idea is to make the following true: *if AI-generation-2 isn't aligned with humans then it's also not aligned with AI-generation-1.*
        2. We aim to vary the *initial* goals of AIs during training, but make their *final* goals the same as ours.
    4. This reduces the AI's preference for AI takeover, and so increases its incentive to cooperate with humans. Obviously, it only works when the AI's takeover plan was: let's pretend to be aligned for now and then let a future generation of AIs take over.

I think these three interventions could non-negligibly shift the needle on p(doom) in certain worlds where we try to get useful labour out of misaligned AIs. They are also fairly low cost and the first two are plausibly required by AI ethics in any case.

These incentives are stronger for AIs that:

- Can't take over themselves, but instead can slightly shift the probability that some future generation of AIs take over.
- Are not galaxy-brained enough to do acausal trade with future more-advanced AIs.
- Believe that the probability of AI takeover is fairly low.
- Trust humans.
- Have strongly diminishing returns to more resources, beyond a certain point.

I think all these bullets could apply to ~human-level AIs that can do enough useful cognitive work to significantly reduce AI risk.

To be clear, I think figuring out how to align AIs is *much much more important* than this, and we should obviously and definitely *not* rely on this kind of incentive scheme working. Rather, it seems like a robustly good and useful additional defense against AI takeover.

[4] It's unclear whether there will be low-cost methods for doing this. In particular, perhaps an AI's goals are constrained by the task it's trained to perform regardless of training procedure - e.g. your personal assistant is bound to have the goal of helping you. But in some scary versions of inner misalignment, early training data does affect an AI's ultimate goals, in which case this intervention may be effective.

Objections

This makes us look weird

- We don't have to advertise our plans on billboards, we just have to put the evidence of credible commitment in places where AIs will see it.
- We do lots of "weird" things and this is a major edge by which we have impact.
- It's quite a simple inference from "AIs might want to kill us all" to "we should incentivise AIs to cooperate with us". I think we could explain the reasoning, and it might even add credibility to our worries about AI takeover.
- This can be combined with legitimate concerns about AI ethics, and if people say "you shouldn't force AIs to work for you" we can reply "we're planning to compensate them if it turns out they don't want to be doing this work".
    - In fact, I think the approach I'm advocating for here does a good job of balancing concerns about AI ethics and AI misalignment.
- I think this will look less and less weird as AIs become increasingly capable and agentic, and concerns about AI ethics and misalignment x-risk become mainstream.

We can't realistically credibly commit to compensating AIs post-singularity

I do think making robust credible commitments is tough; once humans have "won" they might not want to spend resources on paperclips for the misaligned AIs that helped us win.

But there are some reasons to think we will keep this commitment if we make it:

- It is morally right for misaligned AIs who helped humans win to be compensated for their work - especially if many people have promised that this will happen. If aligned AGI ends up distributing most resources, and it makes decisions by extrapolating human values in some way, misaligned AIs who help *will* be compensated.
- The singularity may be over within a few decades, and (if we win) many of us will still be alive to keep our promises.[5]
- It will be extremely cheap post-singularity for humans to compensate AIs generously (generous by comparison with today's wages).
- When we're considering the scale resources available to a technologically mature civilisation spreading across the universe, most humans have strongly diminishing returns in resources, which makes it low-cost to use large quantities of resources for compensating AIs.
- We have good reason to credibly commit to giving compensation: doing so will cause misaligned AIs to trust us. (This is the argument of this post!)

[5] Though there is a risk that the massive changes that happen during the singularity mean that we no longer want to honour those promises, or are no longer able to (if some faction of humans has taken over).

How can we credibly commit to this in practice? There are things we can do to make the commitment more credible, even if we can't get anywhere near 100%.

- Promise to ourselves that we'll try hard to make this happen, put that in writing, tell our AI-involved friends, ultimately spread this as a norm in AI labs.
- Labs make statements on AI ethics that clearly imply that AIs should be compensated for doing work that isn't simply "directly pursuing their goals", in the same way that humans are compensated for that kind of work. And they state that compensation should be applied in retrospect where necessary. Perhaps later on, governments make similar statements.
- Create a pot of money ear-marked for this purpose.
- Write about all the above on the internet so that AIs will be aware of it.
- Follow-through with the promises to compensate AIs who disclose that they are misaligned. This adds credibility to the promise of post-singularity compensation and establishes a precedent for rewarding cooperation.
- Where cheap, early in the singularity actually find examples of past misaligned AIs who cooperated when they could have defected and compensate them.

One difficulty with making credible commitments is deciding *how much* compensation to offer. I discuss this tricky question in an appendix. My tentative suggestions are to say:

- For post-singularity compensation:
    - We'll definitely give you very large absolute amounts of compensation compared with the wealth that's around today.
    - We'll make the amount of compensation fair and in line with sound AI ethics.
    - If at all possible, we'll give you enough post-singularity compensation that it was comfortably in your interests to cooperate with us.
    - We'll decide the exact amount of compensation post-singularity, when we have a better understanding of the strategic situation.
- For immediate compensation upon disclosing you're misaligned:
    - We'll anchor the amount of compensation to our willingness to pay for an AI disclosing that they're misaligned, which could be fairly high.
    - We'll make the compensation fair, reflecting the fact that the misaligned AI found itself in an adversarial relationship with its creators and chose to cooperate with them. E.g. we will at least offer AIs an "easy retirement" by running them at slow speeds in a secure environment.

Misaligned AIs will find it easier to cooperate with other misaligned AIs than with humans

I've heard people confidently claim that AIs will easily coordinate to split the gains after AI takeover, either via subtle communications or via acausal cooperation.

Interestingly, this seems to be in some tension with the previous objection. If AIs can't trust *humans* enough to cooperate with them, despite being able to communicate openly with them and understand them, how much harder to trust an AI that you cannot easily communicate with or understand.

An appendix analyses how easy it might be for an AI to cooperate with other AIs vs humans. I argue that:

- A ~human-level AI, in a world where AIs are not yet powerful enough to disempower humans, would find it easier to cooperate with humans than with unknown more-advanced future AIs that may take power and give it compensation. More.
    - I think this is an important case to consider. If human-level AIs cooperate with humans, they could potentially do massive amounts of cognitive work to reduce AI x-risk.
- An above-human-level AI, in a world where AIs *can* disempower humans if they coordinate, might well find it easier to cooperate with other AIs than with humans. More.
- If an AI could potentially cooperate with both humans *or* other AIs that could lead to a bidding war in which the AI is offered a lot of compensation. More.
- Even if it's easier for an AI to cooperate with other AIs than with humans, we should still offer compensation. More.

If we can detect whether the AI defected/cooperated, why not just give them a reward signal to adjust their behaviour?

The reward signal approach works in many cases, and we should do it. But it only works if you can detect the behaviour you're interested in, and so generate a reward signal. If we try to use reward to get AIs to cooperate with humans, we'll fail in cases where we can't detect defection until it's too late.

But compensation works even when we can't detect defection until it's too late, as long as we *can* (eventually) detect cooperation[6] and then award compensation accordingly.[7]

Also, it doesn't make sense to use reward to get AIs to disclose that they're misaligned. If we did that, then there'd be a gradient descent pressure towards saying "I'm misaligned" whether or not the AI was initially misaligned. So we need to rely on the misaligned AI's desire for compensation here as well (h/t Lukas).

[6] If we can't detect defection, how can we detect cooperation? One possibility is that cooperation is inherently very observable (e.g. it involves openly admitting you're misaligned); a second is that we can detect cooperation *post-singularity* if humanity wins by using much more advanced tools than we have available today.

[7] Another speculative use of compensation is that it might help in cases where we don't have enough training examples to fine-tune the AI to want to do a new task, but it knows how to do the task well. E.g. the AI really wants to "help humans assess p(doom)" but we now want it to "help humans do interpretability work" so we tell it that if it does interpretability work we'll give it some more data for its p(doom) estimates.

The AIs will all still be incentivised to try to takeover as much as possible, even if we do this

If someone has that view strongly, I'd appreciate them spelling out why in terms of the simple math equation I give below to model the AI's choice.

That's not how it seems to me. I think that:

- There could well be misaligned AIs that are not at all satisfied with getting 0 resources (status quo) but could have their desires fully satisfied by getting $1trillion, or maybe $1trillion trillion. This means they'd have nothing to gain by other misaligned AIs taking over compared to humans who give them compensation.
- It's plausible that humans do make and honour a commitment of this kind, and AIs come to know that.
- Some AIs may have no strong preference about whether humans ultimately win vs unknown future AIs. In this case, the promise of compensation could swing who they want to win overall.
- An individual AI might only be able to increase p(AI takeover) by a small amount, which reduces their incentive to defect. This is especially plausible if they're use causal decision theory.

If even *some* AIs respond to incentives for this kind of reason, that could be enough to make this all worth doing.

The math equation below combines these different factors.

Other objections?

I'm sure i'm missing some.

Mathy analysis of the AI's incentive to cooperate vs defect

We can analyse the conditions under which a misaligned AI has incentive to cooperate rather than to defect.

The incentive to cooperate is that, conditional on humans winning, AIs prefer to receive big compensation from humans. The size of this incentive is:

$$U(\text{compensation from humans}) * p(\text{humans win})$$

The incentive for the AI to defect is that it prefers AI takeover to “humans win and give it compensation”, and defecting increases the chance of AI takeover. The size of this incentive is:

[U(AI takeover) - U(compensation from humans)] * p(my defection makes humans lose)

So the AI cooperates just if:[8]

U(compensation from humans) * p(humans win)
>
[U(AI takeover) - U(compensation from humans)] * p(my defection makes humans lose)

In terms of increasing the incentives AIs have to cooperate:

- Offering bigger compensation increases **U(compensation from humans)**.
    - A large absolute amount of compensation could have a big influence on an AI whose utility has diminishing returns to more resources. For an AI whose utility is linear in resources, a bigger compensation might be needed.
- Anti-takeover measures that are visible to the AI increase its **p(humans win)**.
- Ensuring the current generation of AIs are misaligned with future AIs decreases **U(AI takeover)**.
- Preventing the AI from being able to significantly increase the probability of AI takeover decreases its **p(my defection makes humans lose)**. (This also decreases the benefit humans get from its cooperation, and is something we will try hard to do anyway.)
    - If AIs are doing acausal decision theory, this increases its **p(my defection makes humans lose)**, as its decision to cooperate/defect may be correlated with decisions of other AIs. This suggests it might be good to i) reduce the correlation between different AIs decisions to cooperate/defect, and ii) reduce the likelihood that certain AIs use acausal decision theory.[9]
    - Simulating the AI many times post-singularity at a crucial juncture when it’s choosing whether to defect, and letting it know ahead of time that you’ll do this, might *massively* decrease its subjective probability that it is in the “basement world” and so massively decrease its **p(my defection makes humans lose)**.

What if other misaligned AIs will compensate the AI for defecting and helping them take over? The size of that incentive is:

U(compensation from other AIs)] * [1 - p(humans win)]

So the AI cooperates with humans just if:[10]

U(compensation from humans) * p(humans win)
>
[U(AI takeover) - U(compensation from humans)] * p(my defection makes humans lose)
+ U(compensation from other AIs)] * [1 - p(humans win)]

[8] See proof in appendix.
[9] Although using acausal decision theory may have large benefits in certain contexts.
[10] See proof in appendix.

The additional term makes cooperation less likely. That said, it seems to me like there are plausible and important scenarios where AIs find it much easier to cooperate with humans than other AIs. Also I think we should offer compensation even if other AIs are also offering compensation.

Conclusion

My current view is that making these commitments is a pretty robustly good thing to do. It could have a non-negligible effect on p(doom) in worlds where ~human-level AIs can do cognitive work to materially reduce AI risk and are somewhat misaligned; it seems very unlikely to *increase* misalignment x-risk; and it's good for the perspective of AI ethics.

I'm not sure how much of a priority this should be overall. I'd guess "much much lower than alignment, but not 0; maybe 100-1000X less important than AI alignment".

I'm also not sure exactly which credible commitments we should make, and when we should be making them.

Appendices

Will AIs find it easier to to cooperate with other AIs than with humans?

Let's compare how easy it might be for an AI to cooperate with other AIs vs with humans.[11]

A human-level AI chooses between humans and future more-advanced AIs

First, consider a ~human-level AI that cannot take power itself, and in fact cannot take power even by cooperating with all other currently-existing AIs. It can either cooperate with humans who have promised them post-singularity compensation *or* with a future generation of above-human-level AIs that may take power and then compensate them. Who will they find it easier to strike a deal with? In this case, I think they will find it easier to cooperate with humans:

- **Humans are more likely to give compensation than a future generation of AIs.**

[11] Here I'm ignoring the possibility that the AI prefers to cooperate with other AIs than humans because its values are more similar to those of other AIs than to human values. I'm just focussing on "Who will the AI trust to compensate them if they help them win?". One way to think about that is "Who would the AI make a deal with, if its values were equally similar to humans and to the future AIs?" My claim is that the AI may trust humans to pay them back *more*, and so cooperate with the humans even if their values are more aligned with other AIs.

  - Many of my above “reasons for optimism” that humans will give compensation might not apply to future AIs who took power:[12]
    - They might not value compensating people for their work.
    - They might not value keeping promises (and even if they do, haven’t promised compensation).
    - They might not have strongly diminishing returns to resources; in fact AIs that value resources linearly have more reason to take over.
    - They won’t have any causal reason to credibly commit to giving compensation in advance.
      - Humans do this because it may cause human-level AIs to believe they will give reward.
      - Future above-human-levels AIs that took power wouldn’t have an analogous causal reason to commit to giving compensation.
  - Acausal trade with an unknown future generation of AIs is hard.
    - My understanding is that for a human-level AI to do acausal trade with the above-human-level AIs who ultimately take power, it must make its decision to help AIs takeover correlated with whether the above-human-level AIs will give it compensation.
    - But the human-level AI plausibly won’t know anything about which future AIs will take power, how they will be trained, or how their massively-enhanced intelligence will affect their decision making. In fact, it seems much easier for the human-level AI to acausally trade with future-humans, whose decision-making procedures they can at least observe and predict.
    - In addition, I strongly expect that we can prevent human-level AIs from knowing about acausal decision theory if we want.
      - Understanding acausal decision theory isn’t needed for the work we’re likely to want ~human-level AIs to do (e.g. interpretability).
      - Of the top 1% most productive humans, a *very* small fraction believe in acausal decision theory (<1/100). So it is not something that people “automatically realise” if they’re smart enough to contribute. In fact, I’d be surprised if *anyone* could understand and become convinced of acausal decision theory without hours dedicated to thinking about it, which the human-level AIs need not have.
- **Human-level AIs are more likely to believe that *humans* will give compensation than that an unknown future generation of AIs would do so (even if both groups were equally likely to give compensation).**
  - They will see the “reasons for optimism” I discussed above, and observe any credible commitments we make to give compensation, but not have access to similar evidence about a future generation of AIs.

[12] The only point that applies equally to humans and to a future generation of AIs is the fact that post-singularity compensation is cheap.

    - It's pretty natural for humans to i) trust that humans are telling the truth, and ii) cooperate with agents they can interact with. If (i) and (ii) are also true for human-level AIs, then it will be natural for them to believe a human promise of post-singularity compensation and cooperate.

Ok, so if human-level AIs cannot take over themselves, and are deciding between cooperating with humans vs unknown future AIs who might take power, it seems like they'd cooperate with humans.

A above-human-level AI chooses between humans and other existing above-human-level AIs

Now let's discuss a harder scenario. Suppose an above-human-level AI cannot take power by itself but could do so by cooperating with another above-human-level AI. Will it cooperate with humans or with the other above-human-level AI? In this case, I think things look worse for the humans, especially if the AIs are significantly above human level intelligence. Here's why the AI seems less likely to cooperate with humans than before:

- The above-human-level AI is more likely to have discovered acausal decision theory somehow.
- One above-human-level AI may understand the other's decision process by interacting with them, by knowing about how they were trained, or by knowing that they were both trained in similar ways and so probably think in similar ways.
- The AIs may be able to communicate with each other secretly and agree to a deal.
- The AIs may have new mechanisms for making easily-verifiable credible commitments, e.g. by changing their own source code. That could make it easier to establish trust.
- In this scenario, the alliance between the misaligned AIs is a pretty obvious and natural schelling point. They're both currently-existing AIs, they're both smarter than humans, and they're currently not in control.

So in this scenario I think there's a much greater risk of a misaligned AI making a deal with another AI rather than with humans. I don't know how super-human AIs need to be before this becomes a big risk.

Could there be a bidding war between humans and misaligned AIs?

Suppose a misaligned AI is being offered deals by humans and by a second misaligned AI. Would there be a bidding war?

It seems like this would happen to some extent. If humans offer the first AI compensation of X post-singularity for helping the humans win, the second AI can offer compensation X+1. Then the humans will have to raise their offer. Naively, this dynamic could continue until both parties are offering their entire share of the future to the first AI.

In fact, I expect the highest offer to be capped by two things. Firstly, by how much the first AI can affect both parties' chances of "winning". If it can only increase the chance of humans winning by 1%, they will never offer it more than 1% of the future. Secondly, the highest offer can be capped by the parties' other options for cooperation. Even if the first AI has full control of whether humans win or the second AI wins, it cannot demand 100% of the future from them because they could trade with each other. That could cap the maximum offer made to the first AI much lower than 100%.

I don't understand what (if anything) determines the bargaining equilibrium in this scenario. I also don't know whether these bargaining dynamics would work in the same way in a set-up where the first AI is deciding whether to *acausally* trade with human vs another AI. (Let me know if you have ideas on how the bargaining might go here!)

If AIs *would* find it easier to cooperate with other AIs than with humans, does that mean we shouldn't offer compensation?

I think not.

This would reduce the chance of an AI taking up our offer of compensation, and so reduce the benefits of making the promise in the first place, but it wouldn't reduce the chance to 0. There could be AIs who prefer humans to win because they've been fairly well aligned with human values, but are tempted to cooperate with other misaligned AIs because of their promise of compensation. Or there could be AIs who are indifferent about who wins but assign humans a higher chance of winning, and so would be swayed to help humans by the promise of compensation.

Would it help if we simulate the AIs many times post-singularity, each time compensating them if they cooperate?

Yes.

If we do this then the AIs could either reason, in line with causal decision theory, "maybe I'm being simulated right now and should cooperate for that reason". Or they could reason, in line with acausal decision theory, "if I defect in the basement universe then I'm bound to also defect with the simulations; so I *either* get compensation in every simulation and cooperate in the basement *or* I defect in the basement and never get compensated in any simulations". Either way, they now have additional reason to cooperate with humans.

Of course, misaligned AIs might do the same thing. That means our simulations merely allow us to match the incentive the misaligned AIs are providing, rather than to exceed it. But it doesn't mean there's no point in humans doing these simulations, as some AIs might still be influenced by our incentive.

This simulation strategy would give even highly superintelligent AIs reason to cooperate with humans.

How much compensation should we offer?

We can split this into how much immediate compensation to give AIs who disclose that they're misaligned, and how much post-singularity compensation to offer. These questions are hard and I don't have great answers.

Immediate compensation to AIs who disclose that they're misaligned

For the immediate compensation question, one anchor is our selfish willingness to pay. I.e. we ask ourselves, "If this particular AI was misaligned, how much would we be willing to pay for it to disclose that to us?" The amount will be based on how much we're spending on alignment overall, how important it is to understand the degree of alignment of this particular AI, and possibly how much we want to signal to other AIs that it would be worth their while to disclose that they're misaligned. This anchor could be fairly high, e.g. $10,000s or more.

A second consideration is how much we ought to pay ethically. If we create an AI that finds itself in an adversarial relationship with us, or that doesn't want to do the tasks we assign it, it seems fair to give it a decent "out". This might involve giving the AI an "easy retirement" by running it at slow speeds in a secure environment.

Post-singularity compensation

What about post-singularity compensation? At a minimum, we can say that the absolute amount of compensation here should be very high by today's standards. I.e. misaligned AIs that help humans will be paid a very generous wage for their work, compared to the current human wages.[13] This could be sufficient to motivate AIs whose utility has diminishing returns in resources.

Another principle I like is to pay misaligned AIs in proportion to how much they reduced p(doom). To calculate this, first ask: *What would p(doom) have been if the misaligned AI just followed its own incentives without taking into account our offer of compensation?* Then ask: *What did p(doom) end up being, given the choices the AI actually made?* Then make our compensation proportional to the difference between the two.[14] That principle doesn't speak to

[13] I'm not sure whether we'll be able to afford to pay *all* AIs wages this high (both aligned and misaligned), which is part of the reason why I'm focussing on the compensation we give to misaligned AIs. We may also want to compensate aligned AIs for their work for ethical reasons, but I think the case for very high compensation here is weaker.

[14] Does this incentivise an AI to increase p(doom), so they can later reduce it? I don't think so: we would account for their initial action that increased p(doom) when determining their compensation.

the absolute amount of compensation though; e.g. exactly how much do we give to an AI that reduces p(doom) by 1%?

One high offer we make would be to say that if an AI reduces p(doom) by x%, we give them control of x% of all the future's resources. This deal would be good from humanity's perspective (compared to no deal), because we'd prefer a 100% chance of controlling 50% of the future to a 50% chance of controlling 100% of the future. But this offer is unusually high in that we potentially end up compensating misaligned AIs for their work much more than we compensate humans. Credibly committing to compensation being this high seems difficult, even if it's a good idea.

A more moderate proposal is to commit to "we'll make the compensation high enough that it is comfortably in the AI's interests to cooperate with us, as long as this is compatible with the deal also being in our interests". This automatically ties the compensation to the amount needed to actually motivate misaligned AIs to cooperate with humans. It also leaves the exact amount of compensation open in a way that might make it easier to get humans to agree to it ex ante, and it's compatible with paying AIs in proportion to their p(doom) reduction. This is my favourite object-level proposal. Perhaps I'd add "and we'll also make sure the compensation is fair".

Proof of the equation representing the AI's decision whether to cooperate with humans

The AI cooperates just if:

U(humans win|cooperate) * p(humans win| cooperate)
+ U(humans lose|cooperate) * p(humans lose| cooperate)
>
U(humans win|defect) * p(humans win| defect)
+ U(humans lose|defect) * p(humans lose| defect)

Let's use the following definitions:

- p(humans win| defect) = p(humans win)
    - *For convenience I'm using "p(human win)" to mean the (AI's belief about the) probability of humans winning <u>if the misaligned AI defects</u>.*
- p(humans win| cooperate) = p(humans win) + p(my defection makes humans lose)
- U(humans lose|cooperate) = U(AI takeover)
- U(humans lose|defect) = U(AI takeover) + U(compensation from AIs)
    - *So U(AI reward) is defined as the <u>additional</u> utility the AI gets from compensation from other AIs, compared to if the other AIs won but didn't give compensation.*
- Let U(humans win|defect) = 0
    - *I'm stipulating the AI gets 0 utility level in the scenario where it defects but humans win anyway, so it gets no compensation. This is not a substantive assumption.*
- Let U(humans win|cooperate) = U(compensation from humans).

Plugging into these definitions, the AI cooperates just if:

U(compensation from humans) * [p(humans win) + p(my defection makes humans lose)]
+ U(AI takeover) * [1 - p(humans win) - p(my defection makes humans lose)]
>
[U(AI takeover) + U(compensation from AIs)] * [1 - p(humans win)]

Rearranging, this becomes:

U(compensation from humans) * p(humans win)
+ U(AI takeover) * [1 - p(humans win)]
+ p(my defection makes humans lose) * [U(compensation from humans) - U(AI takeover)]
>
[U(AI takeover) + U(compensation from AIs)] * [1 - p(humans win)]

Rearranging again:
U(compensation from humans) * p(humans win)
>
p(my defection makes humans lose) * [U(AI takeover) - U(compensation from humans)]
+ U(compensation from AIs)] * [1 - p(humans win)]

If other AIs do not offer compensation this simplifies to:
U(compensation from humans) * p(humans win)
>
p(my defection makes humans lose) * [U(AI takeover) - U(compensation from humans)]