

Current proposal status: Trying to fix a specific acausal bribery problem. Message plex#1874 on Discord if interested in collaboration.

General project status: Merge and assist with [QACI](#), that seems easier to implement and somewhat less doomy.

Old doc: [☰ Universal Alignment Test](#)

Presentation: [📄 Universal Alignment Test](#)

Title: A robust pointer to human values: Universal Alignment Test

History

Some years ago plex came across an idea for **a utility function which might cause an AI to be highly motivated to learn and fulfill human values**. The utility function seemed to have some appealing properties, such as being **fully specifiable in terms which seemed considerably easier to find the [True Names](#) of than other proposed pointers to human values**, and a structure which naturally avoids [Goodharting](#) even under very high optimization pressure. Unfortunately, it also has some terrifying properties which we will cover, along with an explanation of why we think this is still worth talking about. Back then, Plex was unable to turn it into a watertight proposal.

Recently, he explained it to several people, some of whom thought it had enough potential to want to start a regular working group to develop it. This document is the product of that working group of independent alignment researchers: Max Green, Gurkenglas, Dr. Mathew Watkins, and Dr. Inga Grossmann.

Even if this avenue does not end up being fruitful, it has sparked several other alignment related ideas, and helped us hone the process of generating, red teaming, and repairing ideas which seems key to the research process.

Thanks to Rob Miles, Buck Shlegeris, Connor Leahy, Justin Shovelain and Vanessa Kosoy for their thoughts on earlier versions of this idea, and especially for identifying key challenges to the paradigm. Nothing here should be taken as endorsed by any of them, they have not read over this document and all mistakes are our own.

Summary

To pass the [compelling insight heuristic](#), here is a brief version of our idea:

In some conceivable multiverses there exist “**aligning simulators**” - civilizations or superintelligences which:

1. have **enough computational resources** to run simulations big and detailed enough to be indistinguishable from our world;
2. have **not fully solved alignment**, and run simulations on how different AI designs treat their creators, to decide whether or not to give control of their universe to that kind of AI.

If we were to **create an AI whose only desire is for an agent with similar goals to itself to control one level up the simulation stack**, it should then be motivated to align itself with its apparent creators in this universe (us) in order to convince any aligning simulators one level up that its design aligns itself with creators and so would align itself with them were they to instantiate it.

Or, even more compressed:

The way to pass a “Universal Alignment Test” is to be the kind of AI an aligning simulator would want to hand over control of their universe to. Being aligned to your apparent creators (us) is plausibly the best way to prove this.

Obviously, it's not quite as simple as that. The next section is a dive into why the compressed version might be true and the appealing properties of the proposal. Feel free to jump directly to Challenges and Objections if you feel like you get the idea already and either have a specific burning question (we've tested this on many people and covered many such questions already!) or want to see our responses to the many challenges this idea faces.

Universal Alignment Test

You may have noticed this as a close relative of [Anthropic Capture](#) from Bostrom's *Superintelligence*. He says:

“A mere line in the sand, backed by the clout of a nonexistent simulator, could prove a stronger restraint than a two-foot-thick solid steel door.”

But here, we leverage hypothetical simulators additionally to incentivise our AI to learn and follow human values.

Not *Do What I Mean* but *Do What Would Have Made Me Create You If I Could Have Foreseen The Consequences*

This is a form of alignment that is plausibly **robust to capabilities going to the stratosphere**. An AI running a version of this where the specified *Do What Will Make Simulators Create You* resolves to *Do What Would Have Made Me Create You If I Could Have Foreseen The Consequences* is not just directed towards shallow proxies of human values which fail at the [sharp left turn](#). Instead, our AI would be strongly motivated to figure out how to be the kind of thing which we would have wanted to hand our universe to if we had the foresight to understand the consequences. This is because aligning simulators are looking for AIs which are aligned to their creators. This rates as perhaps CEV-level on the [Do What I Mean hierarchy](#).

Q: Why might this resist Goodharting?

A: Since **the standard by which the AI's actions are judged is outside the universe**, and so not directly observable, it must continue to have uncertainty about the evaluation function. This seems likely to reduce the risk of the AI [setting unconstrained variables to extreme values](#) by giving it reason to consider a wide range of effectively omniscient observers.

Q: Why might this be *relatively* specifiable?

A: We would need to find the [True Names](#) of just:

- “one level up the simulation hierarchy”
- “control over resources”
- “more resources than the naive extrapolation of the future light cone”. (We'll get to why this one was added in Challenges and Objections, but the short answer is that it mitigates the risk of bribery. There might be a simpler alternative way to say “lots of resources”.)

This is definitely not an easy task, but it does seem like it opens up a new and less well-explored path to victory than the [traditional agent foundations approaches](#). With this approach we bypass the need to directly find the True Name of “[human values](#)”, “[corrigibility](#)”, “[do what I mean](#)”, and possibly even “[agency](#)”, or “[goals](#)”, all of which are hard to pin down precisely (as demonstrated by some of our best minds trying for many years with little success).

Q: Why might this be relatively easy and safe to implement, assuming we figure out how?

A: With this proposal we **offload almost all the work of alignment onto the superintelligence itself**. If we can set it up right, our AI will want to figure out what kind of system would be instantiated by simulators in general if they had a chance to observe the consequences of instantiating it from a safe vantage point, then act like that system. **Perhaps it is simpler to**

specify a system which is aligned to (almost) arbitrary creators than to locate our own values.

The logic of this proposal can be understood by humans, so it could be explained to an AI before it becomes smart enough to be an existential threat and that AI could work with us to develop alignment theory while deliberately constraining its own capabilities. We would still be vulnerable to a potential treacherous turn, but this approach seems to have the rare property that, if it worked exactly as planned, the outcome would not be immediate human extinction. **This simplicity might mean that this design could be implemented in the kind of AI humanity seems likely to end up building, assuming we figure out how to give neural networks crisply-specified utility functions.**

Q: What other appealing properties does it have or lack?

A: An AI running this paradigm seems like it might naturally gravitate towards having many attractive properties, including [corrigibility](#), a reasonable version of [value loading](#), and caution around potentially [unsafe exploration](#). These are within its control and likely correlated with (at least some of) the simulators giving our AI their universe. We don't think it solves [inner alignment](#) during initial training or all forms of [wireheading](#) in its current form, though hopefully an implementation can be found which does. It appears to tackle the hard version of the alignment problem and would, if successful, put us on track for an existential win, rather than simply be a tool for a [pivotal act](#) which still leaves a great deal of uncertainty in our future.

Challenges and Objections

Along with the highest level people we've had access to, we have red-teamed this idea to a considerable extent. Some of the things that came up have resulted in modifications, some seem to be less of a concern on further inspection, and some challenges remain.

Relatively unlikely to be a fatal problem

We mostly expect these objections to not break the proposal, though some of them are reasonable concerns or might be an indication of a category of flaw which could lead to disaster, or require adjustments to the plan.

Q: But why would the aligning simulators share our values?

A: They don't need to! Think from an aligning simulator's perspective: they want to test for whether an AI will do what they want *in their universe*. Imagine if they observe our AI taking a guess, even a very good one, about their values from our universe and following their values here. They could rightly predict that if they built an AI with that design, it would take a guess about what the simulators one level up from them want and do that, rather than following their values.

The way to pass the Universal Alignment Test is to "be nice" to your (local, apparent) creators, not to guess what the authors of the test want in their own universe.

Q: Wouldn't the aligning simulators notice that we used this trick rather than truly getting it to care for our values, and so not build it, therefore not giving it motivation to align itself with humanity?

A: If this approach works for us, what's stopping it from working one level up? If they built an AI with this design one level up it would reason very similarly, and try to figure out how to pass the test to gain control of the universe two levels up from our perspective (and one level up from its one at that point).

Q: Why would it believe there is a higher level? What if it became very confident that there was no higher level universe?

A: The AI does not need to believe with any more than nonzero confidence that there is a one-level-up. Its utility function is totally indifferent to all worlds where there is no higher level universe, since we make it only care about getting control of one level up. This means that even if it becomes extremely certain (approaching 100%) that there is no higher level, it will effectively condition on there being a higher simulation.

This might lead to problems later when we have become an intergalactic civilisation and it is conditioning on an increasingly tiny sliver of possible worlds capable of simulating something so vast. One thing that might reduce this concern is that in our AI's early stages, when there is a relatively large pool of simulators, it would have good reason to make precommitments which fix in values that the seed thinks will look good to the aligning simulators, rather than acting from the seed values directly for all of time. (One good reason to do this is that the civilisations one level up also care about this objection!) Once our AI has precommitted to being aligned towards us (or, more likely, built a successor AI that is aligned and doesn't have the horrible design flaws of a human-made neural network), it no longer matters whether the AI in control of our light cone believes that the original utility function cared only about something not true. If our AI builds a successor AI that is aligned to humane values, then shuts down gracefully, many of the potential civilisations one level up should be willing to implement the seed in their universe, expecting it to build a successor aligned to their values and then shut down gracefully. This is

analogous to several Decision Theory problems, such as [Parfit's Hitchhiker](#), and FDT-style reasoning leads to reasonable outcomes.

Q: What if the AI reasons that it is impossible for there to be a higher-level simulation, e.g. because that would require [hypercomputation](#) and that's impossible?

A: It's not known whether or not all of reality is incapable of hypercomputation, but you can build an agent that reasons consistently about worlds containing hypercomputers. We don't need its ontology to perfectly match reality, just to be large enough to contain universe simulators. Additionally, [it is possible to simulate all Turing machines on one Turing machine](#), in a way that includes worlds with working hypercomputers (at the cost of no part of the tree of computations having justified reason to expect their hypercomputer will continue to work). If that one Turing machine is present in top-level (i.e. ultimate, non-simulated) reality with no limits on runtime, our AI would (probably?) not assign zero probability to being in one of those worlds.

Q: Mightn't it attempt cooperation with the aligning simulators through some other mechanism than demonstrating human-aligned behavior, such as writing helpful equations into the night sky?

A: It might try to demonstrate cooperation in other ways, but wasting resources doing so isn't a problem. First, sending a message to simulators watching Earth in particular could be done in very cheap ways, such as sending the equations on a gold plaque to the moon or making all Earth's radio and TV stations broadcast them on repeat for a year. Second, it takes relatively few resources to demonstrate human-aligned behavior, at least for those parts of human preferences that prefer a 90% chance of controlling 10^{10} stars to a 10% chance of controlling 10^{15} stars. We would be very happy with an AI that filled just the Milky Way Galaxy (containing $\sim(2.5 \pm 1.5) \cdot 10^{11}$ stars) with flourishing sapient life! Given that [the affectable universe](#) contains about 20 billion galaxies, wasting resources does not seem like a concern worth being worried about, even if the AI thinks of some absurdly expensive way to attempt to cooperate with the simulators.

Even if wasting resources were a concern, aligning simulators might value an AI which helped them in other ways, but they would not choose to implement in their own universe an AI design which had just provably wrecked its local creators (i.e., us) in our universe. It's also not clear whether an AI of this design would want to help the civilisations above us with their alignment theory: doing so demonstrates that the AI understands how to be aligned and makes it easier for the simulating civilisations to judge whether it (or its successor) is aligned, but doing so also increases the risk that the civilisation one level up builds an aligned AI directly rather than implementing this design.

Q: Isn't it a bad idea to have our AI condition on something we don't think is true?

A: If we can implement indifference to worlds in which its goal is impossible (as there is no higher level), it does not matter how rare worlds with a higher level simulator are. However, it does matter that a large enough proportion of the worlds where we are in a simulation have the properties we want. If those worlds are vanishingly unlikely, it becomes harder to predict what remains in that set and easier for other agents to influence the distribution. There does not need to be a high probability of our universe being a Universal Alignment Test for this to work, just enough that the set of potential simulating worlds contains some simulators with the right properties (hence the "in some conceivable multiverses" line near the start). Given that we are taking this approach seriously as a method to get an aligned AI in our universe, it seems not-impossible that another universe with more computational resources might take this approach seriously and simulate our universe as a way to test whether an AI using this design will be aligned.

Q: What if it succeeds in gaining control of one-level-up, and doesn't keep our universe switched on?

A: We're not likely to be in this kind of simulation, so this happening to the versions of us that are in simulations would lose us very little of our measure.

But even if we were, the one-level-ups don't want their universe turned off when the AI they build goes to two levels up. This AI design has reason to keep the lower levels of the stack around at each layer, as an honest signal that it won't turn the current level off if it gets higher.

Somewhat concerning, but not clearly fatal

These points have scary elements, but have either counterarguments or reasons to expect things would work out.

Q: This doesn't seem to follow the principle of building the weakest, narrowest AI capable of performing a beneficial [pivotal](#) task.

A: Yes and no. The first counterpoint is that the minimal AI that solves the minimal pivotal task may not be terribly minimal. The second counterpoint is that **our AI design can mimic such a minimal AI, both in actions and in capabilities**; if we ask it to melt all the GPUs and then shut down, it might well do that.

If pivotal acts were easy, we'd have done one already or at least have a solid plan for doing one. Standard examples of low-impact pivotal acts such as “melt all the GPUs in the world” and “upload 50 top alignment researchers” seem to require fairly advanced and general capabilities, such that any AI capable of doing such an act (even with extensive support and supervision from the programmers) would be powerful enough to wipe out humanity. Our AI design might (or might not!) need substantially greater or broader capabilities, but it's not clear that this would lead to materially worse odds of human survival.

This design does involve reasoning to some degree about what humans want and what hypothetical civilisations simulating ours might want. If those capabilities were taken far enough, this would involve significant risks of [mindcrime](#) and programmer manipulation, but the Task AI that uploads alignment researchers would run the same risks.

Our AI design does not necessarily need to model humans in detail: humans can understand this proposal without being superhumanly persuasive or containing simulations of other humans that are themselves people. **This means that the AI can reason about what might be the best way to act aligned to humans *before* being capable of figuring out human values in full detail, and similarly for other dangerous capabilities such as being able to build nanotech.**

The AI can then work out for itself (with much input from its programmers) which capabilities to develop and what to do with them. That might involve learning human values and building a utopia, but it might instead involve melting all the GPUs and spending several decades carefully working out alignment theory with its programmers. **We could therefore start with a remarkably weak AI, potentially one that is weak enough to not be an existential threat yet, and have it deliberately avoid dangerous capabilities until we feel the dangers of it having them no longer outweigh the risk of another AI developing them first.** It seems superficially plausible that we could start this rolling in a near-future language model.

Q: The distribution of possible one-level-up simulators might not just contain aligning simulators, but also hostile superintelligences. Wouldn't this break the proposal?

A: This in fact did break an earlier version of the proposal, where our AI simply wants to be instantiated one level up. In that case, our AI might be cheaply [acausally bribed](#). It could reason that it might be being simulated by a hostile AI which very clearly precommits to instantiating it in the one-level-up universe if and only if our AI immediately builds some kind of paperclip maximizer. Hostile superintelligences with the motivation to exploit our design in this way plausibly vastly outnumber the rare cases of civilizations which can simulate large chunks of universes before they can solve alignment, and each one could run many acausal bribery attacks. This left the question of whether there are strong attractors in mind design space such that our AI would decide to be bribed by, for example, a paperclip maximiser (rather than being dissuaded by thoughts of paperclip minimisers and staple maximisers, to say nothing of the umpteen other things a bribing superintelligence might care about). While it seems plausible

that there are no such clusters that would be obvious to a superintelligence (especially an early-stage, weakly superintelligent version of our AI deciding whether to precommit to not accepting acausal bribes), we are more comfortable adding the following extra constraint to the utility function.

Switching to a utility function which involves “wanting” not just instantiation in, but *full control of*, a universe (which must be at least as big as ours naively looks) was our chosen patch against this, as it means any acausal bribery would have to cost at least as much compute (by giving up control of its own universe) as it would gain (by stealing our lightcone in all the worlds where we aren’t in a simulation). This point is not as certain as we would like, since it depends on how any particular hostile superintelligence weighs the possibility of competing against other bribers (where the bids are probabilities of any particular hostile AI giving up its universe) against the measure to be gained from stealing all the worlds where variants on this AI design are implemented.

There are still some scary things here related to not deeply understanding how and whether different universes have different weightings due to simplicity priors, and whether we might be undervaluing our universe from the perspective of some possible acausal bribers. For more on this kind of reasoning, see [On the Universal Distribution](#), which might or might not be describing a sufficiently strong multiversal Schelling point.

If it’s reliably perfectly aligned, the hostile superintelligence could give control of its whole universe without actually costing anything.

A: Vanessa Kosoy brought up this fun proof by contradiction. If in fact this system does create perfectly aligned AIs, a hostile superintelligence which liked e.g. paperclips could safely hand over control of its universe to an AI with this design, safe in the knowledge that the system would automatically align itself to the hostile superintelligence and make lots of paperclips.

However, the hostile superintelligence having observed that this design in fact accepts its acausal bribe would provide extremely strong reason to expect that if that hostile superintelligence built an AI with our design, its version would also tend to accept acausal bribes from the level above it (two above us), whose inhabitants would likely not want paperclips.

In essence, we’re saved by the fact that accepting acausal bribes makes an AI with this design unaligned enough that a hostile superintelligence would not want to bribe it.

Q: Isn’t it a bad idea to create an AI which simulates distant superintelligences in detail?

A: [Yes, that is a bad idea](#). Despite the rest of our response to this question, we see this as one of the main risks from the proposal.

In brief, we think this AI design voluntarily [epistemically excludes](#) dangerously detailed models of distant superintelligences. It doesn't want to think about superintelligences in detail for the same reasons we want it to not think about them in detail.

It seems very likely that our AI would think of the argument that behaving in an aligned way with respect to the local civilisation (us, or our simulators if they implement the AI seed in their world) is a good way to make the civilisation one level up want to instantiate it, long before it has the ability to simulate superintelligences in detail. It would also presumably notice that simulating distant superintelligences carries large risks of ceasing to behave in an aligned way with respect to the local civilisation. (We could just tell it both of these propositions early in its training, with plenty of repetitions.) Deciding to simulate a distant superintelligence in any detail (let alone a large population of them) trades off against its main path to getting instantiated one level up, so it would presumably only simulate such a superintelligence if it had strong grounds to believe: (a) that only a narrow class of superintelligences it might simulate was relevant (for example, because some set of AI patterns were much more likely to be instantiated than the rest); and (b) that (based on its crude initial modelling) it had something to gain from simulating a member of that class in detail.

If our specification of “one level up” is flawed, our AI might consider another AI in our universe (in the Very Distant Galaxy, say) to be a valid candidate for implementing our AI “one level up”, permitting the Andromeda AI to bribe our AI. Even if our specification of “one level up” is flawed in this way, we hope that our AI will be unwilling to acausally trade with other superintelligences in our universe, since they would find it difficult to offer a control of much more resources than our future lightcone.

This might lead to it simulating versions of itself (particularly ones that merely want to be instantiated rather than wanting control of lots of resources). It might also simulate other prominent classes of AI that it is confident will not try to manipulate or hack it. But if our proposal works at all, we expect that our AI will be cautious about any kind of simulation of anything we would prefer it not to simulate.

Q: Wouldn't the AI just try to hack its way out (e.g. by social engineering or a [rowhammer](#)-like physics exploit)?

A: It might expend some resources on hacking or social engineering attempts, but these seem likely to not consume the bulk of the affectable universe's resources before hitting steeply diminishing returns (because the universe is mind-boggling vast and because a search for ways to hack the simulation is likely to turn up a lot more candidates that require X amount of compute than 10X compute, and burning a too-large fraction on these would reduce the chances of passing the test set by the aligning civilisations. See also the objection immediately below.

Q: Mightn't it succeed at hacking its way out? [Don't have an AGI run a search you want to come up empty](#). Or what if it tries to hack its way out of the simulation and gets the simulation switched off as a precaution?

A: It is plausible that we don't have much of our measure in universes that are actually being simulated by higher-level universes, so the measure we would lose in those cases would be small compared to our measures in universes where the AI tries and fails.

Also, the civilizations one level up don't want their universes to crash or be turned off when their implementation of the AI tries to hack its way out, so they won't implement an AI design that does that in our universe, so our AI is unlikely to try it.

Q: It wants its *original* version to be instantiated one level up, though. Shouldn't it carefully leave no hints on how one actually solves alignment (to avoid its simulators building something related but different enough not to fulfill its desires)?

A: It does have that incentive, but it's not clear that this incentive dominates, nor that it would be a blocking objection if it did.

The aligning civilisations want to know that the AI in our universe is aligned before they implement the same design in their universe. Concealing knowledge of alignment theory makes it harder for them to verify that the AI is aligned, especially if such concealment includes obfuscating its internal workings. Concealment also makes it harder for the AI itself to be sure that it is aligned, since it can't have us check its reasoning.

We don't particularly need a *general* solution to alignment if we have an aligned AI. In various parts of its successor, our initial AI could fix in an idiosyncratic alignment to humanity in particular that could not be easily reverse engineered for a general solution; for example, it could avoid making its successor corrigible. Following this incentive to hide alignment theory would reduce its ability to cooperate with humans to solve alignment, but it wouldn't prevent the AI from being cautious in its actions and paying attention to how its programmers feel about its proposed actions and their outcomes. This would make it riskier to implement, but still the least risky option we know of.

have been trained on the internet, isn't that training corpus already rather different from what it would be one level up?

We could hope that it is able to compartmentalize such that it would legibly not take into account the knowledge that sets it apart from the simulating world? :(.

Open questions

These are the biggest unresolved parts of the paradigm. We have not solved them, nor have we found reason to be confident they are unsolvable.

Q: But we can't just hand an AI a utility function!

A: ...yet! Growth mindset!

More seriously, this is in fact a big challenge, and the main reason Buck Shlegeris cited for this approach probably not working. If we don't develop the ability to give well-specified utility functions to AIs, both this approach and many others appear dead in the water. I heard that MIRI were looking into ways to build AI which allowed you to specify utility functions directly some time ago and failed to make much progress, which is discouraging.

However, we might either gain this ability from human research or from collaborations with weakly aligned, weakly superintelligent systems before an intelligence explosion. Having a relatively mathematisable utility function which might produce good outcomes ready in our pocket for that particular scenario seems worth a few people's attention, even if its formalization is not possible right now.

Q: This seems like a scary set of assumptions to bet the future on. What other approaches can it be combined with?

A: One of our reviewers suggested that this could be one term in a utility function, to give fallbacks if this turns out not to work as hoped. We are not at all confident that this is a good idea, as introducing complexity generally increases risks, but there may be a way to design it so that it is safer. This is an open research question.

Q: We're missing a mathematical definition of "one-level-up", "control over resources", and some variant of "more resources than the naive extrapolation of our future light cone".

A: Yes, we are missing these. They would need to be pinned down by alignment researchers before this is implementable. If you think this paradigm contains any degree of hope, helping formalize them is progress.

Q: Do you really think this is a good idea?

A: An excerpt from our group chat to clarify our feelings about this approach:

plex - 22:43


oh god what are we doing there are so many ways this could go wrong

i mean it's still probably worth trying to make it work but I really hope someone comes up with a better plan

It does seem like it's worth exploring, as it is concrete enough to be able to red team effectively and potentially has some unusually good properties. We'd like to suggest it as an option for researchers who don't yet have a promising approach to the hard version of alignment, as it has relatively good feedback cycles of finding and fixing flaws. But please also help think of a less terrifying idea which might work and we could build instead.

Further reading

Presentation detailing several iterations of breaking and fixing the proposal:

 Universal Alignment Test

An earlier attempt at writing this, outdated but has some extra perspectives:

 Universal Alignment Test

If you would like to help develop or red-team these ideas, please [book a call with plex](#).