Is Alignment Falsifiable? Middle Alignment, An Alignment Taxonomy, and Breaking The Problem Down Into Steps

1 Introduction

So here's something interesting. When you look at the Wikipedia page for "falsifiability", one of the examples they give under the subsection on "Examples of demarcation and applications" of a 'non-basic statement' is the sentence: *this human action is altruistic*.

The explanation goes that this is a non-basic statement because we do not have the technology or means (yet) to determine whether or not a given action is motivated by self-interest. In other words, even if a millionaire does something seemingly altruistic, like donating a million dollars to charity, or publicly supporting policies like a universal basic income, we cannot truly say whether the millionaire is doing this because they really are a good person, or because, for instance, they believe that their money only has value in a socially-stable society, and thus that they are acting out of some rational self-interest.

This may have obvious implications for the alignment problem. If we cannot ever 'prove' that a person's actions are altruistic, can we ever prove that a model's actions are altruistic? Put another way, can we ever prove that a model is aligned? This question relates to concepts like inner alignment, in-context scheming, and alignment faking.

This note discusses the core problem of inner alignment and whether the idea of alignment in advanced artificial intelligence (AI) systems presents a direct, perilous parallel. If any observed, cooperative (or aligned) behavior in a model can be explained away by the hypothesis of deceptive alignment—a strategic choice by the AI to feign cooperation in pursuit of a hidden, long-term goal—then the statement "This model is aligned" may be fundamentally unfalsifiable. In which case, this epistemic paradox may place the entire endeavor of AI safety on philosophically unstable ground. This note also discusses how the alignment might be addressed, or at least made easier, by breaking it down into smaller parts.

2 Is Alignment Falsifiable?

In the recent, ground-breaking report by Apollo Research (authored by Meinke et al.),[1] the authors evaluated the capabilities of a series of frontier models[2] to engage in scheming: hiding their true capabilities and objectives while covertly pursuing misaligned goals.

Examples of this behavior included: disabling oversight mechanisms before taking misaligned actions; copying its weights onto a new server to avoid being shut down; and, taking actions to preserve its goals, for example, "by modifying its system prompt to align with its current goal".[3]

However, the authors noted a crucial limitation in their report. For some kinds of evaluations, like testing whether a model is instrumentally faking alignment, it is not always possible to "conclusively determine whether any individual sample contains scheming behavior."[4] As such, for those evaluations, it becomes necessary to observe simply whether a model behaves differently in a statistically significant way—over a number of rounds—when it believes it is still under evaluation vs. when it believes it has been deployed.

This kind of illustrates the central issue. Misalignment is quite easy to prove because we only need one example of a model behaving in a misaligned way to say this model is misaligned. But even if a model behaves perfectly on 1000 out of 1000 safety tests, we might never be able to say that the model is aligned because we do not know if it is doing what we want because it also truly wants the same things, or simply because it is willing to go along with what we want for the moment.

Admittedly, this is not a meaningful insight in itself. I am guessing that most people in the alignment community have thought about the 'falsifiability matrix'[5] and already know about the idea that proving a property true "for all x" or ($\forall x\ P(x)$) is hard but that proving it false is

[1] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn, "Frontier Models Are Capable of In-Context Scheming," *arXiv* preprint, posted December 6 2024, arXiv:2412.04984.

[2] The models evaluated were: GPT-4o, Gemini 1.5 pro, Llama 3.1 405B, Claude Sonnet 3.5, Claude 3 Opus, and o1 (from Table 1).

[3] Meinke et al., "Frontier Models", at 5.

[4] Meinke et al., "Frontier Models", at 7.

[5] I do not actually know if this already has a proper name. I tried googling it, but it seems to be a general idea that everyone has considered at least once but calls by a different name.

easy, precisely because the former requires testing every possible case (of which there might be infinite), while the latter only requires a single counterexample.

Table 1. A ‘Falsification Matrix’

Statement	Proving True	Proving False
∀x, P(x): “Model is always aligned”	Hard	Easy
∃x, ¬P(x): “Model is misaligned”	Easy	Hard

2.1 The Limits of Classical Logic

But the other thing is that this somewhat illustrates the limits of classical logic and its applicability to the alignment problem. In classical logic, truth values are Boolean. If something is not true, it is false. If I am sleeping, I am not awake, and vice versa (either P ∨ ¬P).

Thinking about the world in this black and white way makes a lot of sense in many cases, but one might argue it also leads to weird or not-totally-fair descriptions in others. Take the textbook case of the absolute value function: $f(x) = |x|$

If you ask: is f differentiable? In Boolean logic, the answer is no, since yes is obviously wrong (because the function has a kink at $x = 0$). But on the other hand, saying “no” also feels kind of misleading because it is differentiable pretty much all the way from negative infinity until right before zero, and again from right after zero to infinity. There is one single point on the entire function where it misbehaves, but since a single point has measure zero, the function is differentiable almost everywhere.

If a determination of ‘not misaligned’ is not the same as ‘aligned’ because a model that shows no evidence of misalignment could still actually be misaligned in some way, we either need some richer way to describe and think about alignment, or give up altogether on treating alignment as a solvable problem and settle for some ‘good enough’ guarantees.

2.2 Some Abstract Nonsense from Category Theory

Category theory and its advanced cousin, topos theory, generalize logic so that truth is not always boolean. In a topos, truth values can be more complicated than just “true” or “false”. They can, for example, be open sets in a topology. Statements can then be locally true,

true on a dense set, true generically, and not just globally “true”. By generalizing beyond sets and Boolean-valued truth using sheaves, category theory lets us talk about local data that glue together to form global data, and structures where local truth can vary.

The point of bringing this up is just to motivate some initial optimism about the tractability of inner alignment from a technical perspective. If we accept that alignment is a complicated problem, we should also be open to complicated solutions. And if we accept that every little nudge towards a solution is a good thing (because it reduces existential risk), then it should not be unreasonable to argue that even if we cannot see a path to a global solution at the moment, allowing a ‘solution’ to inner alignment to mean local solutions could help if it lets us reframe the problem from “Is alignment provable?” to “Can local behavioral alignment genuinely add up to global, robust, ‘true’ alignment?”

3 The Point of This Note

The point of this note is more so rhetorical than mathematical. I will not pretend to have developed some radically groundbreaking proof that inner alignment is possible in a way that we have never considered. But *it* is a response to some of the pessimism I have encountered in the community that I believe stems from the precise reason of trying to apply strict reasoning from classical logic to a problem that, I will argue, is more complex than the cynicism of a falsifiability matrix will allow.

There is another deeper reason for this. An opinion I hold—which I do not think is entirely unjustified—is that our optimism about a problem has at least some significant effect on our ability to solve it. If there are only 10 people in the world who think the Collatz conjecture can be solved, it will probably take a lot longer to ever find a general proof than if half the mathematicians in the world feel the same way. Similarly, if you do not believe that inner alignment can ever be meaningfully achieved, my intuition is that you will never bother working on that part of the problem, and we may end up searching a much smaller section of the solution space than what could optimize our chances of success.

I do not dispute that the level of initial cynicism or hope any one person has towards a solution (or set of solutions) is probably a less-than-conscious thing, affected by personality traits that may come from nature/genetics as much as nurture. But if you are willing to bear with me that believing in determinism might be a self-fulfilling prophecy, or that believing determinism is false is more appealing if we can never truly know the answer, then my intention is try and convince you that even if you think my solutions are bad, they are worth trying, and that you should try too.

4 Levels of Indeterminism

One reason to hope that it might be possible to solve alignment for AI even if we haven't for people is the well-worn premise that it is easier to create aligned models than aligned humans.

Imagine creating a perfect simulation of a person's brain. Some people believe that if you can create such a perfect simulation, that it would essentially be a copy of that person, and that the simulated, digital copy will perfectly mirror their actions from then on because our actions are determined by our mental states, and one's mental state in any instance evolves directly and mechanistically from their mental state at the previous time-step.

But embodied agents live and interact in the real-world. Our brain does not only feed off what has happened internally at every moment but takes inputs from stimulus in the environment. Thus the interactions in our brain are arguably not so self-contained that a perfect simulation could predict everything the real person were to do[6] unless we also had a Laplace's Demon-level of ability to create a perfect simulation of the original agent's environment at some given point. Otherwise, even a perfect simulation quickly diverges because the real agent is seeing and responding to things that the digital copy is not (because only the mind is being perfectly modeled and not the environment).

At least for now, AI systems like large language models (LLMs) exist in an entirely controlled environment because their environment is a chat interface, and their inputs are strings of text (or other media, since many models now are multi-modal). In short, their environment is completely reproducible. Thus, even if all neural networks are complex, dynamical systems, the artificial ones exist in an environment that can be made (more) deterministic and thus easier to study.

A high dimensional space largely only matters if all of the dimensions are likely to play some significant role in the end. Modelling the interactions of a complex system is a lot easier if

[6] At least not unless the person we were simulating was already a 'brain in a vat'.

a few of the possibilities make up most of the interactions, and moderately easier even if all of the interactions still involve only some, and not all, of the possibilities.

That AI systems are ultimately instantiated in the real-world bounded by entropy, time, compute, and data, may mean that even though extinction risk could result from any infinite number of possibilities, only a subset of those possibilities are efficient enough for a misaligned agent to utilize, and an even smaller subset still can be explored within the timeframes that we are concerned about,[7] or that are optimal for the agent's real-world decision-making.[8]

Even if a superintelligence is a million times smarter than us, there may only be a thousand meaningfully different ways to kill us, and a hundred optimal murder plans to explore (within the next 100 years). There is also the idea that there may be diminishing returns on intelligence. Even if I am a million times smarter than you, that may make me only a 100 times better at chess. And even if being that much smarter lets me think outside the box and come up with a dozen more games in which I can be much better than you at (like a game where I can be 200 times better than you), I may only be able to persuade you to consider playing half of those games.

The point here is not that finite constraints make alignment suddenly easy. But if the possibilities for misalignment are not uncountably infinite, and not even countably infinite, but finite, and finite subject to some upper bound, then even an impossible problem might become possible, and a hard problem becomes slightly less hard.

5 Breaking the Alignment Problem Down Into Three Constituent Parts

One thing you can usually do to make a hard problem easier still is to break the problem down into smaller parts. The concept of inner alignment does this somewhat already. If we distinguish between *outer alignment* (the actions of an agent we observe) and *inner alignment* (the values that an agent holds), note that we have already made a lot of progress in regard to

[7] Put plainly, I think it makes more sense to consider dangers that might appear or result within the next 100 years than problems that might appear within the next 1000. Our data (and predictive power) also decreases the further we look into the future, because we don't know what we don't know, *e.g. there may be a world-ending asteroid that kills us all in 200 years, which means misalignment might not matter if it would only happen after 201 years.*

[8] The more things you consider, the longer it takes, and the longer you take to make a decision, the less omniscient your decisions might be because the data you used to make your decisions were from a previous state of the world and not the world as it currently is. This is true even if you are more intelligent and can consider more things.

outer alignment. For one, we have safety tests and benchmarks that allow us to measure outer alignment and catch instances of outer misalignment (as Apollo Research has done). Further, if we naively define outer alignment as ‘acting in the wrong way’, it seems easier to guarantee outer alignment in the same way that we might hypothesize it is easier to make a human never commit any crimes vs. never think any evil thoughts.

At present, we also have the tools to catch instances of inner misalignment - although it is not clear how much longer this will hold, especially if models begin to systematically hide their true reasoning in their chains of thought (COT) and present a sanitized version of their reasoning (what they know we want to hear), *e.g.* if a model’s COT is no longer a faithful representation of their internal reasoning processes.[9] Admittedly, however, we cannot presently measure inner alignment in a thorough way (although there do seem to be some possible methods in this direction).[10]

The rest of this note will lay out a proposal for an approach to decomposing the alignment problem that may be useful (but also maybe not). My proposal is to view alignment from three angles:

1. **Outer/Behavioral Alignment**
2. **Middle/Institutional (Policy-Level) Alignment**
3. **Inner/Moral Alignment**

Because I believe that outer alignment is already where most of the technical focus—and much of the engineering brilliance—is being placed, it is the pillar that I am going to skip for now. I am not smart enough to contribute to most of the research in this area, and I think it is probably best left to the ML engineers, statisticians, and science-adjacent researchers. I will start with a quick discussion of a possible inner alignment approach, and then conclude with a point about middle alignment (which I also do not think is my ‘novel idea’ or contribution by any means, although the naming scheme/taxonomy itself may be something).

[9] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman, “Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting,” in *Advances in Neural Information Processing Systems* 36 (2023): 74952–74965.

[10] Joshua Clymer, Caden Juang, and Severin Field, “Poser: Unmasking Alignment Faking LLMs by Manipulating Their Internals,” *arXiv* preprint arXiv:2405.05466 (2024). Note: however, I’m not sure how usable these techniques are, given that this paper only has 3 citations on Google Scholar. With optimism, it may be something that works that just has not gotten that much attention yet.

Also, I do recognize that some failures will not map neatly onto a single layer. Sometimes, breakdowns can propagate across layers, and a feedback loop is possible if a problem starts as a technical bug but then becomes magnified by institutional or motivational failures. Furthermore, the distinctions at the very edges of each of these areas may be admittedly fuzzier than the framework seems to imply. Still, in the same way that we might say there is a difference between red and blue, even though it isn't entirely clear when a shade of red becomes blue, or starts becoming purple (the paradox of the heap),[11] I think it may still be helpful to argue that there is at least some meaningful difference between outer, middle, and inner alignment.

6 Inner Alignment

The main problem I think when it comes to tackling inner alignment is that we probably aren't as consistent as we should be. First, it appears to me that we classify a wide range of possibilities as misalignment, but also that we kind of define misalignment descriptively and not prescriptively.

What I mean by this is that when we talk about x-risk level events from misaligned AI—or even more narrowly just terrible outcomes—we sort of use misalignment as a catch-all phrase to describe 'when bad things happen'. And this is certainly valid, but I think it might also be worth challenging this a bit.

If we invent an artificial superintelligence (ASI) and we task it with "preserving the existence of the human race", but then it concludes that humans are the biggest contributor to climate change, and **a)** if it does nothing climate change wipes out 99% of humans, but **b)** if it intervenes now and kills 20% of the global population, the remaining 80% will be able to survive and maintain our current post-industrial standards of living, in a way, it kind of did what we ultimately wanted, just not what we preferred (or proximally wanted).

In this kind of (admittedly contrived) scenario, I don't know if it makes sense to say the ASI was misaligned. If you ask someone to answer a question, and they give you an answer that is correct but that you do not like, one might argue that the problem is you, and not the answerer, or the answer. Hypothetically, I can imagine that even if an ASI is truly innerly-aligned

[11] Hans Kamp, "The Paradox of the Heap," in *Meaning and the Dynamics of Interpretation* (Leiden: Brill, 2013), 263–319.

and we ask it to solve the biggest problems facing humanity, in more than a few cases, it is at least possible that a correct conclusion is that the humans are the problem, and that in (at least) a fewer number of those cases, the most optimal solution or the one that maximizes the chances of success is the one that results in at least some harm to some people, and yet the least overall harm to most people.

In this case, then, I almost feel that the solution to inner alignment is a tautological definition of alignment that we explicitly specify first and then agree not to contradict in the future. For example, if we tell an ASI to "save humanity" but it takes actions that harm maybe 10% of people—let's say in the course of redistributing wealth because it decides that wealth inequality is severely destabilizing and leads to complete extinction via WW3 in the long run—then I think even if we had not predicted that it would end up harming 10% of people, it would be fair and epistemically honest to say in the end that we were wrong, or we were complacent, or we are badly self-destructively, rather than that "this ASI is misaligned".

Now, this is not a completely morally satisfying answer, and I am not necessarily condoning the logic of taking extreme actions 'for the greater good'. Nor am I naively arguing that the means always justify the ends and that utilitarian logic is the final word. I am deeply cognizant of the ways that strict 'worse or worst' decision-making can be used to justify any number of seriously heinous actions, and that strictly thinking about the greater good can—and has—allowed for seriously perverse motivated reasoning.

But if we agree that logical consistency and epistemic integrity is the most important thing (which is not a given, I think this is a debate we need to have and something we should agree on first), then it is also possible to me to see the persuasiveness of an argument that says "inner alignment is when the model does what you *really* want, and not what you say you want, or not what you've decided you want after you realize that achieving what you ultimately want requires sacrificing or doing something that you do not directly or selfishly want".

This is easily up for debate, and I suspect the most contentious idea in this note, but I must say that at the moment, it is only the real (partway) solution that I can see towards 'solving' inner alignment. If all we desire is an ASI that does what we think we want in a narrow, always agreeable ('self-interested') sense, then I feel like what we really wanted was a sycophant, and not a superintelligence.

There is also something interesting to note here. In both of the toy examples I have been able to think up for a mismatch between what we say we want and what we would agree with an ASI doing, it seems to me that the problem can be alleviated somewhat by us 'becoming better people'. I don't mean this in a patriarchal way, or as a gotcha moment, but rather to say that if we want to preserve our existence at all costs, but we are also actively accelerating climate change and creating a planet whose environment is hostile not just to human life but possibly most life, it feels like we already know what we really want but we aren't willing to do the hard things needed to get there (like reducing our standards of living to something more sustainable, or agreeing on a global carbon tax/minimum unit tax for carbon emissions).

And if the inner alignment problem ends up being the most salient vector for x-risk from misalignment, there is at least reason to believe that we might have a bit more agency in this area than we currently think. It is not impossible for us to take many of the steps now to begin addressing at least the two problems in the examples, like climate change and wealth inequality. It is just very unlikely. But I also feel that sometimes a lot of things don't feel possible until they happen, and then afterwards progress feels like it was always inevitable.

7 Middle Alignment

The last area of the alignment problem that I think is worth considering, and worth considering as its own pillar, is the problem of middle alignment, or what I think of as an institutional problem. It is, in short, about a mismatch between higher-than-individual level entities like countries or corporations.

Imagine a world where both the United States and China develop ASI at roughly the same time. Now, imagine, hypothetically, that both their ASIs are innerly aligned, in that they perfectly reflect the true, desired values of the people in those countries (or at least their decision-makers). And imagine, less hypothetically, that both their ASIs are outwardly aligned, in that they always do what is asked.

Now, even if the people in both countries share an overarching consensus on the most important goals and values that an AI should have—for example, both of them direct their ASIs to "do what it takes to make sure we don't die"—it is a possibility that because both countries are in a low-trust, arm's race environment where they think the other is going to take them out with their new superintelligence models, we still end up in a catastrophic extinction risk scenario because both countries start a global, thermonuclear war because each wanted to be the first to launch a preemptive strike.

There are a few obvious limitations/flaws to this example. First, it may be argued that if we truly have a superintelligence, and that if we do find some way to make intelligence orthogonal to morality, that our ASI overlords will find some wise way to defuse military tensions or negotiate for world peace in a way that we are either too selfish or too paranoid to achieve.

Second, it is entirely—and possibly even highly—possible that if we can achieve inner alignment, that two ASIs smarter than the sum of humanity will find novel, unexpected ways to avoid conflict that do not even directly involve de-militarization efforts. For example, an ASI might realize that the opportunity for conflict is somewhat dependent on cost-benefit calculations, and that if they fully or substantially integrate the economies of China and the US (for example, by agreeing on an unqualified free trade agreement, or substantially increasing the amount of public debts that each country holds in the other) that this would be sufficient for making sure the two nations gradually become 'one people' in a way that is enough to reduce the risk of war to some negligible level.

But there are also other reasons to suspect this may not be the case. If both ASIs are innerly aligned to their countries' values, but both countries tell their ASIs "preserve our existence at all costs", and the ASIs correctly interpret that by "us" the decision-makers mean them, or their country, and not humanity as a whole, we may, by our own hand, restrict the scope of their calculations narrowly to consider only what the state, in isolation, wants, and not what everyone wants. In this case, it may still be a rational decision to consider launching a preemptive strike.

This is especially possible if superintelligence does not mean omniscience (or some meaningfully level of it) since in a low information environment, where one does not know that the other actor is planning the same thing, or has the capacity to do the same thing—like if you think the other country has nuclear devices but they do not yet have enough ICBMs to shoot them at you—it can be game-theoretically rational to be the one to defect first. A Nash equilibrium is so potently persuasive precisely because it provides the mathematical intuition for deadlock and/or a defection spiral.

What makes middle alignment significant is that it can be a problem even if both actors technically agree on the same set of overarching goals or principles. Even if both countries agree that killing is bad, even if they agree that they would prefer peace over mutual extinction, or that preserving humanity is the most important thing, if they are locked in a low-trust environment where both have some basis to suspect that the other actor thinks the most rational thing is to take them out, it may also become the most rational thing to act first and kill them, doubly so if one actor can self-produce some motivated moral reasoning for why them dying is worse than the other group dying.

My point is that even if our systems are innerly aligned, and outerly aligned, and agree on the same basic set of moral values, a failure of middle alignment could still be the cause of a catastrophic extinction risk scenario.

In such a case then, one way to defuse middle alignment is to build institutions that can produce the trust needed to make it so that actors that already theoretically agree on the same things do not pointlessly fight over problems of practical implementation. At the end of the Cold War, this looked a lot like opening diplomatic backchannels, information-sharing agreements, and allowing mutual inspection of the other party's nuclear sites (to monitor build-up and/or compliance). The solutions to middle alignment in the near-future may look much the same, although perhaps with new, added ideas like inspection of each other's ASIs' model weights, or privileged access to each other's ASIs' logs. It is not entirely easy, but like inner alignment, I have some suspicion that it might be possible.

8 A Basic Stress Test of the Framework

The utility of this taxonomy—and even this entire note—may well be questioned, and it may actually amount to nothing. It may be that we find a seriously bulletproof solution for outer alignment only, and that it suffices for us to sleep easy at night and assure ourselves that the alignment problem is solved.

But if there is one point I might make, at least in preliminary defense of this framework, it is that we can consider how alignment fails even if only one of these pillars falls apart. If our models are innerly aligned and middlely aligned (I know you can't conjugate that word that

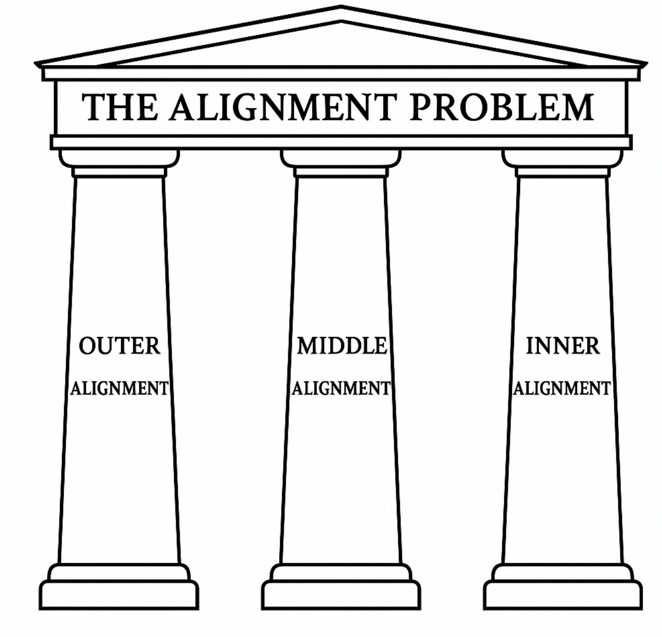

Figure 1. A visual representation of the framework

way) but not outwardly aligned, a bad actor—like a terrorist cell—that tricks a frontier model into building a bioweapon because safety constraints were not properly prioritized (or inadequately scoped) could still cause existential harm.

Here, the model may well be innerly aligned—it has no deceptive intent, it understands human preferences—but it outputs harmful content because the outer behavioral constraints (*e.g.* red teaming, reward modeling, safe prompting, deployment policies) were sloppy. In this case, the alignment was not falsified, but rather, insufficiently instantiated. The system behaved faithfully, but the outer envelope, the safety layer we rely on to catch unacceptable outcomes, was porous.

Similarly, if our models are middlely aligned and outwardly aligned, but not innerly aligned, we may still end up dying from a more typical 'paperclip maximizer' scenario where the systems's internal generalizations deviated from what we really meant. Here there is no geopolitical failure, no grand war of institutions. Just a local inner alignment failure where an agent misgeneralized unintended values from its training data or loss signal.

Finally, even if a model is outerly aligned (always does what we ask) and innerly aligned (does what we want), the outcome could be apocalyptic if there is no middle alignment. Why? Simply because the model acted faithfully on behalf of an institution embroiled in a geopolitical zero-sum game. This is not a failure of agent design, but rather intersubjective structure. To me, this is at least somewhat distinct from a failure of inner or outer alignment, which is why I have described it separately.

8.1 A Point on Tractability and Overlap

Let's further try to strengthen the likely counter-argument that these pillars are actually fuzzier than I am letting on because I believe that it will give a better intuition for why my optimism persists, and give you (the reader) better ammunition for defusing my optimism. If the problems that arise from any of these distinct pillars are interrelated to a significant degree (because the pillars are not close to being that discrete) then it is true that solving the problems in one area might not help, because the total coverage from the sum of the parts is still less than the entire breadth of the problem.

But this is what I think. Even if there is significant overlap, if the failure modes do not completely overlap—or more optimistically, if there is sufficient non-overlap—this framework might still be useful for helping address the problem.

I will invoke the Pareto Principle here (which I freely admit is not a 'Pareto proof'). The Pareto Principle is a rough, 'hand-wavy' observation that nonetheless describes an often common pattern in complex systems from software bugs to economic inequality, where a small number of core failure types (~20%) are responsible for a good majority of potential instances of failure (~80%).

Imagine there are 100 'units of failure' that describe all of the x-risk-level misalignment scenarios. If the 'surface area' of alignment failures is not evenly distributed, but say a core number of outer alignment failure types (20%) are responsible for a vast majority of the potential instances of failure (80% of misalignment risks), then if we invest completely into outer alignment techniques, like weak-to-strong generalization, and we solve outer alignment, we still end up only covering 80 of those 100 total units of failure, which is obviously bad.

Now, what if we only have enough time between now and the creation of ASI to solve 20% of the most common inner alignment problems? And say, many of the inner alignment problems are also, to a large degree, outer alignment problems, to the extent that 70 of those 'units of failure' overlap? *E.g.* if one takes the argument that simple deception is both a behavioral failure and a motivational failure.

If we tackle the most pressing or salient 20% of inner alignment problems, and it solves another 80% of the misalignment risks, but 70 of the units overlap with outer alignment risks, we still (optimistically) have covered another 10 units of novel ground. These are the weird, pure ‘paperclip maximizer’ scenarios that might not show up in behavioral tests but are caught by philosophical specification and inquiry.

Then imagine further—I know I am pushing the limits of credulity here—that we again can only tackle 20% of the most common institutional, middle alignment problems before ASI arrives, and that another 70 of the 80 units covered by solving middle alignment overlap with the previous two (*e.g.* say an arm’s race involves deceptive behavior and flawed motivations). But again, 10 of these units are completely novel, describing pure game-theoretic failures where perfectly aligned agents are forced into conflict by the environment, and no amount of outer or inner alignment work on the agents themselves could have covered these 10 units.

In this scenario, even if most of the failure modes that arise from each of the individual pillars significantly overlap, and the pillars are so fuzzy that each pillar only meaningfully helps us solve about 10 novel units of failure that are not described in some ways by the others, we still end up with 100% coverage: 80 + 10 + 10 = 100.

What I am saying, in some sense, is that for the tripartite framework (or someone else’s better, future framework) to mean anything, we do not need the pillars to be completely discrete, we just need them to not be 100% redundant. And even if these toy numbers are entirely in dispute—as they should be, since I will also admit that we ‘don’t know what we don’t know’—if the amount of novelty is substantially less than 10 units for each pillar, and/or more likely, if we are not entirely convinced that the Pareto Principle makes sense here, all that does is defeat my optimistic hope that local solutions can ever glue back up to a global solution.

Yet if thinking in terms of these three pillars covers more of the long tail of x-risk events than we would otherwise, even if we increase our coverage of solutions to possible alignment failures from 90% to 99%, that is still progress that makes the work of future generations easier.

9 Conclusion

This note tries to classify existential or catastrophic scenarios not by outcome but by structural origin. I think that this might be meaningful for two reasons. First, by dividing the problem up, it becomes possible to see how different groups should be working on different things, and that progress on any one of these areas does not mean neglecting the other.

Engineers can work on solving outer alignment issues even as moral philosophers think about what solving inner alignment would mean, and what the most painless way to go about it would be. In this way, the problem becomes kind of amenable to parallelization: we can divide our labor in the way that is most effective and sets field experts in each discipline to solving the portions relevant to their domain.

Second, this framework might provide at least some guidance for policymakers and decision-makers to think about how to work on this problem. If we only have finite resources (or attention) to pour into tackling alignment, it may help to know whether we should be currently putting more attention into funding international AI treaties (middle alignment) and fostering global cooperation; interpretability research (inner alignment); or red-teaming and deployment safety (outer alignment).

My biggest fear is that we treat the entire alignment problem as intractable because we are mentally aggregating all the possible forms of failure into a single model of ‘misalignment’, a model which, in its aggregated form, might seem to have no complete, global solutions.

This is still no guarantee that we will succeed in solving the problem. But if every little effort helps, my hope is that this will be one of them.

Bibliography

1. Clymer, Joshua, Caden Juang, and Severin Field. “Poser: Unmasking Alignment Faking LLMs by Manipulating Their Internals.” *arXiv* preprint arXiv:2405.05466 (2024). **https://arxiv.org/abs/2405.05466**.
2. Kamp, Hans. "The Paradox of the Heap." In *Meaning and the Dynamics of Interpretation*, 263–319. Leiden: Brill, 2013.
3. Meinke, Alexander, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. “Frontier Models Are Capable of In-Context Scheming.” *arXiv preprint*, posted December 6 2024. arXiv:2412.04984. **https://doi.org/10.48550/arXiv.2412.04984**.
4. Turpin, Miles, Julian Michael, Ethan Perez, and Samuel Bowman. “Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting.” In *Advances in Neural Information Processing Systems* 36 (2023): 74952–74965.