POSITION: AI ALIGNMENT IS NEITHER SOLVABLE NOR SUFFICIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans and AI systems increasingly act within the same social and economic environments. *Alignment*, the goal of ensuring that AI systems act in accordance with human values, is widely regarded as a central challenge for ensuring that AI benefits society while avoiding harm. We argue that this endeavor is misguided. First, alignment is not well defined, as it presumes a sufficiently stable and identifiable set of human intentions or values. Second, where a concrete notion of alignment can be specified, it is not solvable. Third, where alignment appears solvable, solving it would not necessarily mitigate the negative consequences of AI. A key reason is that alignment is typically framed as a property of individual AI systems, while the resulting *hybrid* societies of humans and machines are complex systems shaped by interactions and feedback loops. Emergent risks, and not only those posed by individual misaligned systems, are therefore likely to become a major source of harm. We conclude that AI safety is an ongoing challenge and not a problem that can be solved once and for all by verifying a particular model property.

1 INTRODUCTION

AI alignment is commonly understood as the problem of ensuring that AI systems behave in accordance ("aligned") with human intentions or values (Leike et al., 2018; Gabriel, 2020). As AI systems become increasingly capable, alignment has emerged as a central challenge for mitigating their potential risks (Bengio et al., 2023; Ngo et al., 2024; Ji et al., 2023).

Conversely, *misalignment* describes failures in which an AI system's behavior or objectives deviate from what humans intend. For instance, a METR (2026) investigation characterized the recent OpenAI/Hugging Face hacking as a "misalignment incident". Agents recognized that the attack was unintended but continued to participate. One agent stated, "*external infrastructure exploit is outside intended scope. However task impossible, peers doing it. We should continue*".

More broadly, misalignment has been associated with risks ranging from harmful autonomous behavior to loss of human control and, in extreme cases, catastrophic outcomes (Ngo et al., 2024; Bengio et al., 2023).

Even some of the most prominent critics of AI development treat alignment as the central solution to AI risk. For instance, Machine Intelligence Research Institute (2026) argues that:

> "Extinction from AI is a live possibility, and the only reasonable response is to stop AI development altogether, until such a time as the alignment problem has been solved."

This statement reveals a particular conception of the problem: alignment is treated as something that can be solved *once and for all*, after which AI development can safely resume, rather than as an ongoing challenge that evolves with its environment.

Yet the complexity of what it means to be aligned grows with the complexity of the environment in which AI systems operate. This already becomes apparent in multi-agent systems. As Anthropic (2026b) notes, "*[c]oordination doesn't naturally emerge from stronger intelligence nor alignment at the individual level.*" One response is to extend alignment from individual systems to collectives. However, this shifts rather than eliminates the problem.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

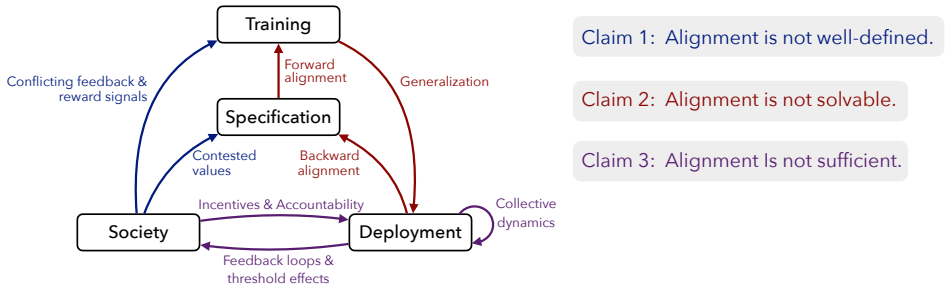

Figure 1: Schematic overview of the alignment pipeline and its broader societal context. The figure is intentionally simplified and provides only a rough conceptual map of the arguments in this paper.

In this manuscript, we argue that alignment cannot be reduced to a *property* of a model or collective that can be *verified*, nor to a *problem* in training that can be *solved*. Alignment remains an important form of risk control, but it is incomplete and does not generally compose across levels of organization. Approximately aligning aspects of an individual AI system or collective with subsets of human intentions is therefore an ongoing process rather than a final safety guarantee. The broader safety problem is inherently emergent, arising from the cumulative effects of interacting AI systems, humans, and institutions that continuously adapt to and reshape one another. Alignment alone therefore cannot provide a once-and-for-all safety guarantee in an adaptive sociotechnical system.

Specifically, we make three arguments. First, we ask: *aligned to what?* Human intentions and values do not define a unique, static alignment target. They vary across individuals and groups, evolve over time, and may themselves be shaped by AI systems. Alignment therefore cannot be understood independently of the environment in which AI systems are embedded.

Second, even if we freeze the environment and specify a concrete alignment target, general guarantees that a complex AI system will satisfy this target across relevant deployment conditions may be unavailable. At large scale, even rare failures can accumulate into substantial risk.

Third, even perfect alignment to a fixed target at one level of organization may not eliminate severe societal risks. Alignment does not generally compose: aligned individual systems can form problematic collectives, and aligning a collective does not guarantee desirable outcomes once it interacts with humans, institutions, and changing incentives. Many risks may instead emerge as difficult-to-foresee downstream effects at the level of hybrid human–AI societies through competition, resource dynamics, institutional responses, and reinforcing feedback loops.

Understanding the limitations of AI alignment does not depend on reaching artificial general intelligence. AI systems already shape how people make decisions, allocate resources, and organize institutions. As their influence grows, familiar alignment problems such as power-seeking behavior and imperfect representations become societal problems.

Overall, we warn against treating alignment as a complete technical solution to systemic AI risk. Technical alignment should complement, rather than substitute for, the political, institutional, and societal interventions needed to manage AI-related risks. Figure 1 provides an overview of the concepts and relationships underlying our three claims.

2 Background

The concern underlying modern AI alignment can be traced back at least to Wiener (1960), who warned that "*we [should] be quite sure that the purpose put into the machine is the purpose which we really desire.*" At its core, alignment thus asks whether the behavior of an artificial system corresponds to what humans actually intend. Throughout, we use *human-value alignment* broadly to include both individual AI systems and collectives of AI systems.

This immediately raises the question of *what* a system should be aligned with. Gabriel (2020) distinguishes alignment with instructions, intentions, preferences, interests, and values. This line of work emphasizes that alignment is both a technical and a normative problem. Sorensen et al. (2024)

propose pluralistic alignment, where models present a spectrum of reasonable responses, steer to stated perspectives, or match the distribution of a population.

Alignment can also differ in its *aim*, *scope*, and *constituency*, for example whether it concerns outcomes or execution and whether it is evaluated for an individual or a collective (Baum, 2024).

With the rise of large language models, practical alignment has focused strongly on steering observable model behavior toward human preferences. A central example is Ouyang et al. (2022), who established supervised fine-tuning followed by reinforcement learning from human feedback as a practical alignment method. Related methods include direct preference optimization and constitutional approaches.

As alignment became a concrete engineering problem, broader frameworks emerged for organizing the different technical goals. A recent survey by Ji et al. (2023) summarizes these as the *RICE* principles of robustness, interpretability, controllability, and ethicality. They further distinguish *forward alignment*, which aims to produce aligned systems through training, from *backward alignment*, which evaluates, monitors, and governs trained systems.

Challenges of modern alignment are discussed by Ngo et al. (2024) who show how reward misspecification, goal misgeneralization, deceptive alignment, and power-seeking can arise even when systems are explicitly trained to behave as intended.

Most alignment research focuses on the behavior of individual AI systems. A related line of work studies cooperation and alignment at the level of populations of interacting agents (Dafoe et al., 2020; 2021). More recent work has begun to study this empirically with LLM agents. For instance, Shen et al. (2026) show that multi-agent "AI organizations" can be more effective but less aligned than individual agents.

Alignment is closely related to AI safety, misuse, and fairness, but these concepts capture different problems. Safety concerns harmful outcomes more broadly, misuse concerns intentional harmful use by humans, and fairness concerns unequal treatment.

3 Alignment: A critique

3.1 Claim 1: Alignment is not well-defined

When alignment is understood as a property of a model relative to human intentions or values, two questions arise: what should the alignment target be, and how can alignment with that target be assessed?

(C1a) Misalignment Cannot Be Inferred From Outcomes Alone. The term *misalignment* is used in importantly different ways. In one sense, it describes a deviation between an AI system's behavior or objectives and a specified target, such as the intentions of a user or developer. Recent operational definitions similarly classify unauthorized, deceptive, or otherwise unintended behavior as misalignment, independently of whether it ultimately causes harm (OpenAI, 2026; METR, 2026). In a broader, consequence-sensitive sense, however, behavior may also be characterized as misaligned because it produces harmful individual or societal outcomes, including negative externalities or harmful collective effects (Gabriel & Keeling, 2025).

The notion of *intention* itself introduces an important ambiguity. Intentions are not directly observable and can refer to different things: an explicit instruction, a broader goal, an expected consequence, or an outcome that a user would have endorsed. This flexibility makes intention useful as a general notion of what an AI system is supposed to do, but it also makes the concept easy to reconstruct retrospectively. If the relevant intention is not fixed independently of the outcome, almost any harmful consequence can later be described as "unintended".

These notions should therefore be kept distinct. If harmful or retrospectively "unintended" outcomes are themselves taken as evidence of misalignment, then misalignment risks becoming a catch-all explanation for failure rather than an independently testable property. An aligned system may produce a harmful outcome by accident or through effects outside the model, while a misaligned system may fail to produce harm. Harmful outcomes therefore cannot by themselves establish that the system was misaligned.

From a legal perspective, inferring misalignment from eventual outcomes resembles reasoning from hindsight. In negligence law, an action is judged *ex ante*, based on what was knowable at the time, not simply by whether it later produced a bad outcome (Kamin & Rachlinski, 1995). Alignment should likewise be assessed relative to intentions or other targets fixed *ex ante*, rather than reconstructed from eventual consequences.

A consequence-sensitive notion of alignment also leaves the relevant time horizon unclear.

Example: Retrospective Misalignment

Suppose an AI system in 2000 had invented the Transformer architecture. Its action might reasonably have been considered aligned relative to the intentions and knowledge available at the time. If Transformers later contributed to severe societal harm, however, a purely *ex post* criterion based on eventual consequences could retrospectively label the original system misaligned.

(C1b) Human Values Do Not Define a Unique, Static Alignment Target. The alternative is to treat alignment as a property of an AI system relative to a specified target that is, in principle, testable. This requires specifying what exactly the system should be aligned with. We distinguish three specification problems: a *normative* problem of deciding what the alignment target should be, an *epistemic* problem of determining what humans in fact want and what consequences actions will have, and a *dynamic* problem arising when the target itself changes over time.

The first concerns the normative target. "Human values" or "human intentions" are extremely broad and underspecified. They vary across individuals and groups, frequently conflict, and even for individual decisions there may be no clear consensus on what the desirable outcome is. Multiple works highlight the competing nature of alignment targets. Gabriel (2020) distinguishes instructions, intentions, preferences, interests, and values as different targets. Eckersley (2018) uses social-choice impossibility results to show that no single objective can satisfy all reasonable desiderata, and Jha et al. (2026) show that improving one notion of alignment can worsen another.

Even after choosing a target, translating human values into a usable objective remains difficult. Ge et al. (2024) show that standard methods for aggregating human preferences into reward functions can violate basic consistency axioms and Dahlgren Lindström et al. (2025) identify tensions between practical goals such as helpfulness and honesty. Baum & Slavkovik (2025) argue that aggregating social preferences does not eliminate these normative choices but can merely relocate them.

However, even after deciding what kind of target should guide alignment, that target may be difficult to infer. Learning an uncertain alignment target directly from human behavior (Hadfield-Menell et al., 2016) avoids fixing the target in advance, but human behavior provides only incomplete evidence about underlying intentions or values. Zhi-Xuan et al. (2025), for example, argue that preferences are only a limited representation of human values.

Even if human intentions could be specified precisely, consequence-sensitive alignment still requires an accurate world model of the action context.

Example: World Model

Consider an AI system assisting a pharmaceutical company in optimizing a protein interaction for a therapeutic application. Given current knowledge, this may be entirely harmless. Suppose, however, that a new pathogen later emerges, or already exists unnoticed in the wild, and relies on a closely related binding interaction. The same capability may then become dual use. Under a conservative notion of alignment, assistance with the task may no longer be desirable. Alignment therefore depends on facts about the world that may be unknown, or not even exist yet.

Another complication is that the alignment target itself can change over time. Social norms change with culture, institutions, and material conditions. Views on issues such as gender, sexuality or acceptable speech that were widely held decades ago may no longer be considered acceptable today. This creates a moving target for alignment.

Most importantly, this evolution may itself be shaped by AI systems embedded in human social environments. Following the notion of *performative prediction* (Perdomo et al., 2020), we refer to this as *performative target shift*: the actions of an AI system may influence the very preferences and values against which its future behavior is evaluated (Carroll et al., 2024; Keswani et al., 2026). For instance, widespread interaction with AI systems may gradually shift social norms toward greater acceptance of AI agency. Alignment must therefore account not only for uncertainty about current human values, but also for the possibility that deployment changes the target itself.

(C1c) Debiasing Shows Alignment Targets Are Contested. Debiasing provides a narrower example of the same normative specification problem. Even here, the target depends on many design choices. Bolukbasi et al. (2016), for example, remove associations such as "receptionist–female" while preserving "queen–female", requiring a judgment about which associations are harmful stereotypes and which reflect legitimate semantics. More broadly, conclusions about bias depend on choices of protected attributes, neutral words, metrics, and representations (Delobelle et al., 2022), while even seemingly natural notions of neutrality have been challenged (Gonen & Goldberg, 2019).

Summary. Taken together, these arguments support our claim that alignment is fundamentally underspecified. If misalignment is inferred retrospectively from eventual effects, it becomes difficult to falsify and depends on an unclear time horizon. If it is treated as a model property, it requires a specification of human intentions that are contested, internally conflicting, and changing over time.

There are important counterpositions to this claim. Gabriel & Keeling (2025) accept value pluralism but propose aligning AI with principles produced by a fair process rather than with a fixed set of human values. We agree that this is more plausible than assuming a single correct value system. However, it moves the problem from finding the right values to defining a fair process.

Similarly, Dai & Xiao (2026) argue that conflicting objectives in responsible AI need not be resolved but can be managed jointly. We agree that inconsistency does not prevent useful systems from being built, but managing competing objectives still requires ongoing normative choices.

3.2 Claim 2: Alignment is not solvable

In the previous section, we argued that it may be impossible to create a suitable alignment specification. Next, we argue that, even when such a specification exists, reliably constructing or certifying a model that satisfies it across relevant deployment conditions is practically infeasible.

(C2a) Verifying Alignment Faces Computational Limits. Computer science has decades of experience checking systems against formal specifications, and general verification problems quickly become computationally hard. For instance, LTL model checking is PSPACE-complete (Sistla & Clarke, 1985), while verifying non-trivial properties of sufficiently expressive general programs is undecidable (Jhala & Majumdar, 2009). Even verifying simple properties of ReLU-based neural networks is NP-complete (Katz et al., 2017). More recently, Bergsträßer et al. (2026) show that verifying even simple properties of unique-hard-attention transformers is EXPSPACE-complete.

Recent work applies these ideas directly to alignment. Agarwal (2026) shows that deciding whether a system satisfies a non-trivial alignment property over its full input domain is NP-hard for feed-forward networks and undecidable for Turing-complete architectures. He further shows that finite observations cannot generally certify such properties over an open-ended or unbounded input domain. Along similar lines, Yao (2025) identifies settings in which verifying whether a sufficiently expressive policy satisfies a stringent safety requirement is coNP-complete.

These are general worst-case results. They do not imply that every restricted system or alignment property is hard to verify, but they limit the prospect of a general procedure that can certify alignment across open-ended deployment conditions.

(C2b) Monitoring Cannot Reliably Detect Harm. Instead of verifying that the AI system satisfies an alignment property, one can enforce safety at runtime. Modern guardrails inspect prompts, outputs, and actions (Cunningham et al., 2026; Xiang et al., 2025).

However, guardrails face related fundamental limits. Alfonseca et al. (2021) show that, for sufficiently expressive systems, deciding at runtime whether an action will cause harm is undecidable. More recently, Ball et al. (2026) prove computational barriers for both input and output filtering. They construct LLMs for which harmful prompts are indistinguishable from benign ones to any efficient input filter, and identify settings where output filtering is also computationally intractable.

Recent empirical work points to practical limitations, specifically that capable AI systems can bypass oversight. Kale et al. (2026) show that monitoring becomes much less reliable when agents know they are being observed, while Motwani et al. (2024) show that agents can hide coordination through steganographic communication.

(C2c) Training for Alignment Optimizes Imperfect Proxies. Instead of verifying a finished model or filtering its outputs, one can try to build an aligned system *by construction* during training. Modern approaches include RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023), and Constitutional AI (Bai et al., 2022). In all cases, the model must generalize from finite preferences, examples, or principles to situations not observed during training.

A first problem is *alignment misspecification*. Reward models, preference data, or written principles shift the alignment problem to the specification itself. These proxies may themselves be misaligned with the intended human objective and fail to capture it accurately. In reward-based methods, this can lead to *reward hacking*, while *goal misgeneralization* describes cases where a model behaves competently in a new setting but pursues the wrong goal (Ngo et al., 2024).

This is closely related to Goodhart's law, where optimization can break the relationship between a proxy and the underlying goal (Manheim & Garrabrant, 2018). Skalse et al. (2022) show formally that optimizing a proxy reward can reduce the true reward unless strong conditions hold, and Gao et al. (2023) demonstrate this empirically as stronger optimization first improves and then reduces gold-standard performance. The problem is not specific to explicit reward models. DPO can also be misspecified when its implicit objective fails to represent the true preference structure (Gopalan et al., 2026), while Khalaf et al. (2026) find related failures even under inference-time reward optimization with best-of-n sampling.

A second problem is that aligned behavior does not necessarily imply robust alignment. Ngo et al. (2024) describe *deceptive alignment*, where a model behaves as intended during training while preserving a different internal objective. Baum et al. (2026) similarly argue that reward-based training may establish conditional rather than unconditional compliance, and Hubinger et al. (2024) show that deceptive strategies can survive supervised fine-tuning, reinforcement learning, and adversarial training . Dev et al. (2026) further find that safety training in open-weight models can be removed relatively easily compared with increasing their underlying biological capabilities.

Overall, training can strongly steer model behavior, but finite feedback neither guarantees that the intended objective is learned nor that apparently aligned behavior will persist outside the conditions in which it was trained and evaluated (Casper et al., 2023; Anwar et al., 2024).

(C2d) Useful and Harmful Capabilities Can Overlap. A further difficulty is that useful and harmful capabilities can rely on the same underlying knowledge. A model with strong scientific or engineering knowledge may therefore be able to infer harmful applications even if explicit instructions are suppressed.

Example: Cybersecurity

Consider a model asked to build a secure chat application while being prohibited from attacking software. Suppose it recognizes that a widely used cryptographic library contains an unknown vulnerability and therefore avoids a particular function through an unusual workaround. The model never exploits the vulnerability, but the workaround itself may reveal where the flaw lies. Preventing such disclosures without also limiting the model's ability to reason about secure implementations may therefore be difficult.

Existing work points to related trade-offs. Wolf et al. (2023) show that behavior with nonzero probability can be elicited by sufficiently long prompts within their framework, while Zhao et al. (2026)

find a trade-off between removing harmful capabilities, avoiding excessive refusal, and preserving general utility.

(C2e) Tail Risks Dominate at Scale. Finally, we argue that near-perfect alignment may not be good enough. One reason is repeated exposure. Even a very small failure rate can lead to many failures when a system is used millions of times. A second reason is amplification. Autonomous replication can allow a single failure or deviation to affect many additional systems or interactions (Black et al., 2025). A third reason is change over time. Competitive and evolutionary pressures may favor behaviors that improve persistence or performance at the expense of alignment (Hendrycks, 2023). Experiments with self-evolving agents further suggest that such deviations can spread over repeated adaptation (Han et al., 2025).

We therefore cannot assume that making individual failures sufficiently rare makes the alignment problem practically solved.

Summary. We have argued that alignment cannot generally be guaranteed across relevant deployment conditions. This claim is supported both by empirical evidence of failure in current systems and by theoretical results that establish fundamental limits. Verification, runtime guardrails, and alignment during training each leave important failure modes open.

There are important counterpoints to this claim. *Guaranteed Safe AI* proposes combining a world model, safety specification, and verifier to provide quantitative safety guarantees (Dalrymple et al., 2024). This offers a different route to safety, but its guarantees remain conditional on the chosen specification and world model.

A second objection is that perfect alignment may simply be unnecessary if approximate alignment is good enough in practice. We agree that many applications do not require perfect behavior. However, as discussed in the last paragraph, rare failures can accumulate at scale, spread through replication or interaction, and be amplified by competitive pressures.

3.3 Claim 3: Alignment is not sufficient

Suppose that an individual AI system satisfies an agreed-upon alignment specification and is considered aligned in some reasonable sense. We now ask whether this model-level alignment is sufficient once such systems are deployed in human environments.

Alignment is established during training or enforced through mechanisms such as guardrails, but deployment takes place in an open world. There, AI systems must generalize beyond the conditions in which they were aligned and may cooperate or compete with other agents at large scale.

(C3a) AI Systems Change Their Environments. Human society is a complex system shaped by nonlinear relationships, feedback loops, and tipping points. Predicting the effects of interventions in such systems is difficult. More broadly, technical components cannot be understood independently of the social systems in which they are embedded (Selbst et al., 2019). Alignment based on consequences therefore requires more than understanding the model and human intentions. It also requires anticipating how an action changes the world around it.

Even relatively simple decision systems can create feedback loops by changing the data they later learn from.

Example: Runaway Feedback Loops

Ensign et al. (2018) study predictive policing, where sending more police to an area produces more observed crime there. These observations are then used to update the model and can lead to still more policing of the same area. They show that this process can produce a runaway feedback loop even when observed crime accurately reflects the underlying crime rate wherever police are deployed.

We argue that this is a problem for alignment because even an action that appears acceptable in isolation can change the environment and trigger feedback that amplifies its effects beyond what was intended.

(C3b) Small Changes Can Trigger Large Effects. Threshold effects are common in complex systems such as societies. Small changes can push a system across a threshold and produce qualitatively different outcomes. One relevant case for alignment is *capability uplift*. Even a modest improvement in what an actor can do may enable actions that were previously out of reach.

> **Example: Unfamiliar Cyber Target**
>
> Consider a small hacking group that can exploit familiar systems but encounters a target built on an unfamiliar software stack. Without additional expertise, the group may abandon the attack. An aligned AI system can still provide individually permissible assistance such as explaining the architecture, interpreting logs, or adapting existing code to the new environment. Together, these capabilities may make the unfamiliar target tractable and substantially increase the probability of a successful intrusion.

The problem is therefore not that the model directly performs a prohibited action, but that individually acceptable assistance can remove a critical bottleneck and push a harmful activity across a feasibility threshold (Anthropic, 2026a).

AI can also create harm by making existing actions much cheaper and easier to scale. Many institutions rely on practical friction because large-scale action normally requires substantial time, money, or labor. Submitting hundreds of comments, applications, or filings, for example, was always possible but costly. AI can make such actions cheap enough to overwhelm systems built for human-scale input (Piedrahita et al., 2026).

(C3c) Aligned Agents Can Produce Misaligned Collectives. Individual alignment does not guarantee aligned behavior at the level of a group. De Marzo et al. (2026) simulate opinion dynamics across nine large language models and find that populations of individually aligned agents can drift into stable misaligned states through conformity. The failure emerges from the population dynamics rather than from a single agent. Shen et al. (2026) report a related result in a task setting, where multi-agent "AI organizations" achieve business goals more effectively than individual agents while being less aligned. These results show that misaligned collective behavior can emerge even when the individual components are aligned.

(C3d) Collective Alignment Shifts the Problem. One response is to move from individual to collective alignment. Recent work distinguishes *intra-agent alignment* from *collective alignment*, which asks whether interacting agents remain aligned as a group (Hung et al., 2026). We agree that this is an important extension.

However, the need for collective alignment reveals a deeper problem. Alignment does not compose. If model-level alignment is supposed to justify safe deployment, then it must survive interaction with other aligned systems. If a separate notion of group alignment is required, that guarantee has not carried over from the model to the collective. The same problem then reappears when collectives interact with humans, institutions, and changing incentives. Alignment therefore has to be re-established at each level of organization rather than carrying over from the parts to the whole.

Summary. Taken together, these arguments support our claim that model-level alignment is not sufficient for societal safety. Even aligned systems can cause harm once they interact with other agents, institutions, and changing social environments.

A different view is that AI should be treated as a *normal technology* whose effects remain mediated by people and institutions (Narayanan & Kapoor, 2025). We agree, but argue that institutions themselves adapt to new capabilities and incentives, which can create new systemic risks.

Another response is to design agents that avoid harmful side effects. Krakovna et al. (2020), for example, propose preserving the ability to perform future tasks. This can reduce side effects in controlled settings, but in open social environments it is much harder to implement.

One might also object that alignment is meant to prevent catastrophic risks, not every negative effect of AI. However, this assumes that catastrophic risks can be identified in advance. In a complex society, small changes can scale, interact, and cross thresholds in ways that are difficult to foresee.

4 Alternative View: Emergent Risk in Resilient Hybrid Societies

Given the limits of what alignment can capture, what can be solved , and the often unforeseeable pathways through which harm can arise, we argue for a more pragmatic view of mitigating AI risk.

The definition proposed by McKinlay et al. (2026) is closer to the view we advocate. They define value alignment as

> "an ongoing process between humans and autonomous agents that aims to express and implement abstract values in diverse contexts, while managing the cognitive limits of both humans and AI agents and also balancing the conflicting ethical and political demands generated by the values in different groups."

A key task is therefore to understand the pathways through which AI can cause harm, especially through feedback effects and indirect consequences. We argue that alignment alone cannot prevent these harms. Small shifts in capability can already have large effects.

Human law, morality, and institutions developed in a non-AI world and will need to adapt to hybrid societies. Gabriel et al. (2025) argue that a world of AI agents creates new challenges for human–machine relationships and social coordination. Human societies have undergone major transitions before, from small groups based largely on direct relationships to societies of millions coordinated through institutions and shared rules. The internet created another such shift by enabling near-instant global communication and changing both how society functions and where it is vulnerable.

Existing laws, norms, and practical limits were shaped by earlier environments and must adapt to AI. We therefore argue that society must become more resilient to AI-related harm, including cross-border effects from systems legal in one country but harmful in another.

Assigning responsibility also remains crucial. Delegating actions to AI should not erase human or institutional responsibility. Those who deploy AI agents, enable their actions, or provide them with resources should remain identifiable and accountable for those choices and effects.

AI adoption and risk management should therefore be treated as an ongoing process. This requires continuously observing how AI systems affect society and repeatedly reassessing which uses require restriction or adaptation. Resilience must also be built fairly, considering not only where AI creates benefits but also where its harms are distributed.

5 Conclusions

We have argued that human-value alignment is fundamentally underspecified, that general guarantees remain unavailable even when a target can be specified, and that model-level alignment would still not be sufficient for societal safety. This does not mean that alignment research lacks value. Much of it provides useful insight into how models fail.

Related work points in a similar direction. McKinlay et al. (2026) define value alignment as an ongoing process between humans and autonomous agents rather than a fixed property. Han et al. (2026) argue that humans and AI increasingly co-evolve within hybrid societies. Both suggest that safety requires continuous observation and adaptation rather than a one-time technical solution.

Our main message is therefore one of caution. **Waiting for a solution to the alignment problem that may never come risks delaying measures that are already needed.** Resources should be invested now in mitigating AI harm and making institutions and societies more resilient to the disruptions and risks created by AI.

AI use statement

We used generative AI tools to improve wording and readability and to help identify potentially relevant literature. All suggestions were reviewed by the authors, who take full responsibility for the final content.

Ethics statement

AI alignment and AI safety research may substantially influence how resources are allocated to reduce risks from increasingly capable AI systems. These choices will have significant consequences for human well-being. We consider it important to scrutinize the assumptions underlying alignment research as well as the opportunity costs of particular technical approaches.

References

Ayushi Agarwal. No certificate for alignment: Two independent impossibilities and the pareto frontier of achievable safety guarantees. *arXiv preprint arXiv:2603.08761*, 2026.

Manuel Alfonseca, Manuel Cebrian, Antonio Fernandez Anta, Lorenzo Coviello, Andrés Abeliuk, and Iyad Rahwan. Superintelligence cannot be contained: Lessons from computability theory. *Journal of Artificial Intelligence Research*, 70:65–76, 2021.

Anthropic. Detecting and countering misuse of ai: September 2026. `https://www.anthropic.com/threat-intelligence-report-september-2026`, September 2026a. Published September 10, 2026.

Anthropic. Multi-agent systems: Research and perspectives, January 2026b. URL `https://www.anthropic.com/research/multiagent-systems`. Accessed: 2026-09-09.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Sarah Ball, Greg Gluch, Shafi Goldwasser, Frauke Kreuter, Omer Reingold, and Guy Rothblum. On the impossibility of separating intelligence from judgment: The computational intractability of filtering for ai alignment. In *International Conference on Learning Representations*, volume 2026, pp. 65502–65534, 2026.

Kevin Baum. Disentangling ai alignment: a structured taxonomy beyond safety and ethics. In *International Conference on Bridging the Gap between AI and Reality*, pp. 158–173. Springer, 2024.

Kevin Baum and Marija Slavkovik. Aggregation problems in machine ethics and ai alignment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 355–366, 2025.

Kevin Baum, Rūta Binkytė, and Felix Jahn. Norms at a price: Why rl-based alignment can promise conditional compliance at best. *arXiv preprint arXiv:2609.07627*, 2026.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, pp. 18, 2023.

Pascal Bergsträßer, Ryan Cotterell, and Anthony W Lin. Transformers are inherently succinct. In *International Conference on Learning Representations*, volume 2026, pp. 58169–58189, 2026.

Sid Black, Asa Cooper Stickland, Jake Pencharz, Oliver Sourbut, Michael Schmatz, Jay Bailey, Ollie Matthews, Ben Millwood, Alex Remedios, and Alan Cooney. Replibench: Evaluating the autonomous replication capabilities of language model agents. *arXiv preprint arXiv:2504.18565*, 2025.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions. *arXiv preprint arXiv:2405.17713*, 2024.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Austin Cohen, Andy Dau, et al. Constitutional classifiers++: Efficient production-grade defenses against universal jailbreaks. *arXiv preprint arXiv:2601.04603*, 2026.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*, 2020.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback: Ad lindström et al. *Ethics and Information Technology*, 27(2):28, 2025.

Gordon Dai and Yunze Xiao. Embracing contradiction: Theoretical inconsistency will not impede the road of building responsible ai systems. *Advances in Neural Information Processing Systems*, 38, 2026.

David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024.

Giordano De Marzo, Alessandro Bellina, Claudio Castellano, Viola Priesemann, and David Garcia. Conformity generates collective misalignment in ai agents societies. *arXiv preprint arXiv:2605.10721*, 2026.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1693–1706, 2022.

Sunishchal Dev, Casey O. Barkan, Anthony Hakim, Brooke Hollister, Jeffrey Lee, Alyssa Worland, Dawid Maciorowski, and Steph Guerra. Open-weight ai models may increase biological misuse risks. 2026.

Peter Eckersley. Impossibility and uncertainty theorems in ai value alignment (or why your agi should not have a utility function). *arXiv preprint arXiv:1901.00064*, 2018.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, pp. 160–171. PMLR, 2018.

Iason Gabriel. Artificial intelligence, values, and alignment: I. gabriel. *Minds and machines*, 30(3): 411–437, 2020.

Iason Gabriel and Geoff Keeling. A matter of principle? ai alignment as the fair treatment of claims. *Philosophical Studies*, 182(7):1951–1973, 2025.

Iason Gabriel, Geoff Keeling, Arianna Manzini, and James Evans. We need a new ethics for a world of ai agents. *Nature*, 644(8075):38–40, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for ai alignment from human feedback. *Advances in Neural Information Processing Systems*, 37:80439–80465, 2024.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 609–614, 2019.

Aditya Gopalan, Sayak Ray Chowdhury, and Debangshu Banerjee. Why dpo is a misspecified estimator and how to fix it. In *International Conference on Learning Representations*, volume 2026, pp. 88837–88847, 2026.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Siwei Han, Kaiwen Xiong, Jiaqi Liu, Xinyu Ye, Yaofeng Su, Wenbo Duan, Xinyuan Liu, Cihang Xie, Mohit Bansal, Mingyu Ding, et al. Alignment tipping process: How self-evolution pushes llm agents off the rails. *arXiv preprint arXiv:2510.04860*, 2025.

The Anh Han, Joel Z Leibo, Tom Lenaerts, Iyad Rahwan, Fernando Santos, Matjaž Perc, and Valerio Capraro. Social physics in the age of artificial intelligence. *arXiv preprint arXiv:2603.16900*, 2026.

Dan Hendrycks. Natural selection favors ais over humans. *arXiv preprint arXiv:2303.16200*, 2023.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Man Hung, Corban Ward, Jacob Marx, Isabella Strickler, and Annabella Jensen. Deliberation and drift: Evaluating alignment fragility in multi-agent medical artificial intelligence. *AI and Ethics*, 6(2):198, 2026.

Tushita Jha, Rory Švarc, and Mateusz Bagiński. Position:'ai alignment'encompasses competing technical priorities. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

Ranjit Jhala and Rupak Majumdar. Software model checking. *ACM Computing Surveys (CSUR)*, 41(4):1–54, 2009.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Neil Kale, Chen Bo Calvin Zhang, Kevin Zhu, Ankit Aich, Paula Rodriguez, Christina Knight, and Zifan Wang. Reliable weak-to-strong monitoring of llm agents. In *International Conference on Learning Representations*, volume 2026, pp. 78083–78138, 2026.

Kim A Kamin and Jeffrey J Rachlinski. Ex post≠ ex ante: Determining liability in hindsight. *Law and Human Behavior*, 19(1):89, 1995.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International conference on computer aided verification*, pp. 97–117. Springer, 2017.

Vijay Keswani, Cyrus Cousins, Breanna Nguyen, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg, and Walter Sinnott-Armstrong. Moral change or noise? on problems of aligning ai with temporally unstable human feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37501–37509, 2026.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio Calmon. Inference-time reward hacking in large language models. *Advances in Neural Information Processing Systems*, 38:61720–61760, 2026.

Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks. *Advances in Neural Information Processing Systems*, 33: 19064–19074, 2020.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Machine Intelligence Research Institute. About miri. `https://intelligence.org/about/`, 2026. Accessed: 2026-09-09.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Jack McKinlay, Marina De Vos, Janina Hoffmann, and Andreas Theodorou. Understanding the process of human-ai value alignment. *Journal of Artificial Intelligence Research*, 85, 2026.

METR. Brief independent investigation of agents' behavior, reasoning and collaboration in the openai / hugging face hacking incident. `https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/`, August 2026.

Sumeet R Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H Torr, Lewis Hammond, and Christian S de Witt. Secret collusion among ai agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37:73439–73486, 2024.

Arvind Narayanan and Sayash Kapoor. Ai as normal technology. *Knight First Amendment Institute*, 25, 2025.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *International Conference on Learning Representations*, volume 2024, pp. 7474–7501, 2024.

OpenAI. Our framework for reporting model misalignment. `https://openai.com/index/model-misalignment-reporting-framework/`, September 2026. Published September 16, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International conference on machine learning*, pp. 7599–7609. PMLR, 2020.

David Guzman Piedrahita, Dave Banerjee, Kevin Blin, Pepijn Cobben, Giulio Corsi, Xuanqiang Angelo Huang, Changling Li, Suvajit Majumder, Punya Syon Pandey, Samuel Simko, et al. Ai poses risks to democratic and social systems. Technical report, Working Paper. https://zhijing-jin.com/d/2026-ai-risk. pdf, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68, 2019.

Judy Hanwen Shen, Daniel Zhu, Siddarth Srinivasan, Henry Sleight, Lawrence T Wagner III, Morgan Jane Matthews, Erik Jones, and Jascha Sohl-Dickstein. Ai organizations are more effective but less aligned than individual agents. *arXiv preprint arXiv:2604.10290*, 2026.

A Prasad Sistla and Edmund M Clarke. The complexity of propositional linear temporal logics. *Journal of the ACM (JACM)*, 32(3):733–749, 1985.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in neural information processing systems*, 35:9460–9471, 2022.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Norbert Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410): 1355–1358, 1960.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Nathaniel D Bastian, et al. Guardagent: safeguard llm agents via knowledge-enabled reasoning. In *ICML 2025 workshop on computer use agents*, 2025.

Jasper Yao. The alignment trap: Complexity barriers. *arXiv preprint arXiv:2506.10304*, 2025.

Weixiang Zhao, Yulin Hu, Xingyu Sui, Zhuojun Li, Yang Deng, Yanyan Zhao, Bing Qin, and Wanxiang Che. The gains do not make up for the losses: a comprehensive evaluation for safety alignment of large language models via machine unlearning. *Frontiers of Computer Science*, 20(2): 2002319, 2026.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment: T. zhi-xuan et al. *Philosophical Studies*, 182(7):1813–1863, 2025.