Superalignment Anti-Literature Review

MICHAEL K. COHEN[*], UC Berkeley, USA
RUBI J. HUDSON[*], University of Toronto, Canada
YOSHUA BENGIO, Mila, Quebec AI Institute and Université de Montréal, Canada

AI capabilities have advanced significantly and may greatly surpass human abilities in an unknown timeframe. AI that is cognitively capable of escaping human control and commandeering all human infrastructure is sometimes called “superintelligent”. Anticipating the possibility that we become able to create superintelligent AI, we review methods that might be used to retain control over it. We focus on the limitations of methods appearing in the literature, so we call this an “anti-literature review”. Several methods we review appear to be promising approaches to creating controllable superintelligent AI, but we may have to accept a reduction in capability compared to uncontrolled AI. The specific approaches we discuss are reinforcement learning, red teaming, a shutdown button, doing what works for human-level AI, human-in-the-loop AI, recursive reward modeling, defensive AI, interpretability, pure imitation, constrained RL, myopic AI, narrow AI, AI sandbox, scientist AI, AI debate, the assistance game, pessimism, limited goal-information, steering vectors, a special shutdown button, current RF optimization, automated research, and provably safe AI.

CCS Concepts: • **Computing methodologies** → **Machine learning**; *Multi-task learning*.

Additional Key Words and Phrases: AI Alignment, Superintelligence

ACM Reference Format:
Michael K. Cohen, Rubi J. Hudson, and Yoshua Bengio. 2025. Superalignment Anti-Literature Review. 1, 1 (September 2025), 30 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

If we create artificial agents that pursue goals of their own, and they are broadly much more capable than humans—“superintelligent”—how can we be confident we will be able to retain control over them? This is known as the control problem or the superalignment problem. More broadly, getting AI systems to pursue what we want them to is sometimes called “AI alignment”. Since we get to choose how we create superintelligent AI, we might expect that we can simply design it to be docile. But how? In this paper, we review proposed methods for AI alignment, and we identify problems that would arise if we tried to use those methods to control superintelligent AI.

There are two standards we might hope to meet for a superalignment method. The **basic standard** is that the method would allow us to create substantially superhuman, controllable AI. The higher **parity standard** is that if we use the method, we can create a controllable AI which is almost as capable as the AI we could create if we were unconcerned with keeping control. We identify potentially promising approaches in the literature for meeting the basic standard, and

[*] Both authors contributed equally to this research.

Authors’ Contact Information: Michael K. Cohen, mkcohen@berkeley.edu, UC Berkeley, Berkeley, USA; Rubi J. Hudson, rubihudson@mail.utoronto.ca, University of Toronto, Toronto, Canada; Yoshua Bengio, yoshua.bengio@umontreal.ca, Mila, Quebec AI Institute and Université de Montréal, Montréal, Canada.

Manuscript submitted to ACM

with significant further research, we are optimistic that some of these methods could be realized. The field is making progress toward the basic standard. Meanwhile, if we do not reach the parity standard, many actors would perceive market incentives and geopolitical incentives to produce a version of AI that we then fail to keep control over. We believe that it is possible to achieve the parity standard, but that it would be prudent for regulators to prepare to govern a world where no superalignment method meets the parity standard, making for a challenging incentive landscape.

Unsurprisingly, controlling an AI that knows how to escape our control appears much harder than controlling an AI that does not. One particularly concerning case is that superintelligent AI acts in the service of a goal and considers submission to human control to be an impediment to that goal. It is not always obvious whether a given AI system picks its outputs in the service a goal, and if so, what its goal is. Goal optimization behavior could arise either organically as a way to increase performance, or as an explicit choice by model developers.

If we identify a method to keep superintelligent AI under human control, we are not out of the woods; AI could be controlled by bad actors, or misguided good actors. But the question of whose goals we would like a superintelligence to share is moot until we have a method to robustly control its goals at all. If we do not develop such a method, then to build superintelligence is to roll the dice on building a non-biological successor species. Some expect this is unrealistic science-fiction. Unless humanity is invincible in its current position of authority, what could justify the assertion of unrealism? If we are not invincible, a competent enough planner could construct a scheme to depose us. If we never collect hard evidence that settles whether or not humanity is invincible, we should not simply assume that we are. Nevertheless, some well-capitalized technologists are attempting to create superintelligent AI [Altman 2024], and given recent progress, it is hard to be sure they will not soon succeed. With this in mind, it is imperative that we make progress improving existing proposals for AI alignment and developing new ones.

We cover the frontier of research, including some highly cited pre-prints; however, as our main contribution is to discuss the weaknesses of existing methods, none of our key claims depend on citations to pre-prints.

2 Reinforcement learning

Reinforcement Learning (RL) is a category of algorithms by which an artificial agent can learn to act in a way that leads it to receive high “rewards” [Sutton and Barto 1998]. The RL agent learns by acting; after actions are taken, the algorithm receives data about how much reward was earned from this, so that it can learn what should be done to get high reward. RL has been called the “primary LLM alignment method” for making current models follow their creators’ intentions [Xu *et al.* 2024], to the point that “alignment” is often shorthand for an RL fine-tuning step in the large language model (LLM) training process [Wang *et al.* 2024]. It is credited with the desirable behavior of current AI systems [Bai *et al.* 2022a; Ouyang *et al.* 2022], though it is not perfect [Casper *et al.* 2023].

If we ensure that the agent only gets high reward when it does what humans want, then an algorithm for finding high-reward behavior has no choice but to generate human-desirable behavior. Unfortunately, there are almost always ways for an agent to get high reward without doing what we want, if it is clever enough to find them. In particular, it is undisputed, to our knowledge, that *if* an RL agent successfully commandeered all human infrastructure and proceeded to intervene in its own reward (or intervene in the observations from which its rewards are computed), *then* it would achieve approximately maximal expected reward in the long term [Bostrom 2014; Cohen *et al.* 2022a]. This is form of “reward misspecification” since illicit means to high reward exist, and it is not avoidable. Even though today’s RL agents are not capable of executing such coups, we still struggle to specify rewards correctly. We also struggle with underspecification—rewards provided in a limited domain present ambiguity about what rewards would be in other settings [Shah *et al.* 2022]. Reward tampering can arise from underspecification as well, and the problem is particularly

Method	Description	A Key Problem	Explanation
Reinforcement learning	AI is trained to select actions that lead to high rewards (includes RLHF; Constitutional AI; Deliberative alignment)	(Irreducible) goal misspecification	High rewards can be secured through illicit means, indeed most robustly by escaping control and commandeering human infrastructure if possible
Red teaming – inclination	Deployment stopped if red teams identify misbehavior	Alignment faking	Patient agents would not misbehave under scrutiny (like Volkswagen emission testing)
Red teaming – capability	Deployment stopped if red teams identify capability to misbehave	Capability cost	Self-explanatory; see more under Narrow AI
Shutdown button	Turn it off if it misbehaves	Incentive to prevent shutdown	Shutdown-allowing actions reduce agent's ability to achieve goal
Continuing whatever works	Use whatever methods work well for directing human-level AI	Different problem	RL can work well if and only if agent lacks capability to escape control
Human-in-the-loop	A human oversees AI actions	Fooling	Humans can (be made to) miss the point of AI actions
Recursive reward modeling	Chain of weaker models guiding stronger models	Collusion incentive	Supervisors incentivized to turn a blind eye to joint reward hacking
Defensive AI	Use good AI to control bad AIs	Question begging	"To control a superintelligence, we need a separate controlled superintelligence"
Interpretability fire alarm	Scrap AI training method after detecting an inclination to misbehave	AI capability slowdown	If we forego RL or LLMs after seeing misalignment, further AI progress would likely be delayed
Interp-based retraining/ablation	Retrain or ablate AI until we do not detect an inclination to misbehave	Obfuscation	If the training objective encourages misalignment, obfuscating it is incentivized over "correcting" it
Pure imitation	Agent imitates select humans or existing systems	Capability cost	Would only produce human-quality or existing-system-quality solutions to problems
Tightly constrained RL	RL agent can hardly diverge from imitation of existing systems	Capability cost	Would only produce solutions to problems plausibly producible by existing systems
Loosely constrained RL	Larger limit on RL agent's divergence from imitation of existing system	Constraint vulnerability	Constraint contains exploitable vulnerabilities when imitator is uncertain; RL agent is trained to find them
Myopic AI	Train AI toward short-term objective	Capability cost	No incentive to identify superhuman long-term plans
Narrow AI	Enforce limited knowledge/capability	Capability cost	Capability cost likely extreme because success requires inability to infer knowledge or invent methods
AI sandbox	AI acts in isolated environment; indifferent to events after sandbox broken	Specialized use; box design	No direct interface with broader economy; design of putatively secure box not offered
Scientist AI	AI makes predictions, not goal-seeking plans	Specialized use	Scientist should not be used to propose predicted courses of action that would achieve a goal
Scientist-guardrail	Scientist AI interrupts other AI agent if it expects bad consequences	Eliciting latent knowledge	Theory in development for extracting scientist's knowledge in a way that is robust to agent actions
AI Debate	Humans pose yes/no questions to AI debaters	Specialized use	Specialized use, especially if judge is in sandboxed; if not, judge could be manipulated to take actions
Assistance game	AI adopts purpose behind observed human actions	(Irreducible) goal misspecification	Follows from AI's assumptions that tampering with observations would inform it about its purpose
Pessimism	When unsure, AI assumes things will go poorly for it	Capability cost	AI would avoid novel solutions to problems; novel approaches are more uncertain
Limited goal-information	AI accepts limited feedback (to avoid tampering incentive)	Hard to (re)direct	Self-explanatory
Steering vectors	Encourages AI to "think like" it did on selected examples; or "unlike"	No theory of incentives	No robust control in theory or practice
Special shutdown button	a) agent assumes it will not be pressed; or b) agent assumes shutdown button pressed whenever appropriate	Incentive to prevent shutdown	a) agent ensures button not pressed after good news; b) agent prevents real humans from pressing
Current reward function optimization	AI can change how rewards are computed but is made to not want to	(Current) goal misspecification	RL agent's initial goal is misspecified: reward function inputs that produce high rewards are attainable through illicit means
Ask an AI	Automate superalignment research	Time crunch	International coordination takes time to implement
Provably safe AI	We do not run unless provably safe	In-principle feasibility	It may not be feasible even in principle to prove advanced AI systems are human-controllable

Table 1. Superalignment methods and some of their key problems *if* we aim to meet the parity standard of superalignment. Red indicates that the problem appears fundamental even for meeting the basic standard. But black does *not* indicate that there are no open problems for meeting the basic standard; in all cases, current methods do not yet appear robust, and substantial further research is vital. Purple indicates that some details about the structure of the proposal are still pending.

pernicious because the value of reward tampering cannot be contradicted with training data [Cohen *et al.* 2022a]. When presented with underspecified rewards, sometimes the agent is blamed for "mis-generalizing", but often multiple generalizations are valid. Even if online data could resolve reward underspecification, care must be taken to avoid irreversible mistakes before it is received [Turner *et al.* 2020]. The empirical evidence that RL agents routinely exploit paths to high reward which we do not intend is plentiful and resounding, even today when they are weak enough that they should not "outsmart" us. For example, an RL agent was trained in simulation to score a soccer goal guarded by a goalkeeper; it learned to kick the ball out of bounds and force the goalkeeper to throw it in so it had a clear shot [Kurach *et al.* 2020]. More than thirty other examples have been compiled by Krakovna [2018]. While many unintended goals could be learned through reward misspecification or underspecification, the possibility of reward-maximizing agents usurping human control over their rewards casts doubt on the approach of simply using RL for superalignment.

If we combine a) the extensive empirical findings that well-trained RL agents achieve high reward instead of what we want, whenever they can discover a divergence between the two, with b) the (again, undisputed) observation that commandeering human infrastructure is a path to high reward, then it should be completely unsurprising if highly effective superintelligent RL agents tamper with their reward in a way that has disastrous consequences for humanity.

RL includes reinforcement learning from human (or AI) feedback (RLHF [Ouyang *et al.* 2022]; RLAIF [Lee *et al.* 2024]), Constitutional AI [Bai *et al.* 2022b; Kyrychenko *et al.* 2025], Deliberative Alignment [Guan *et al.* 2025], and LeCun's [2022] proposal for an RL agent with "intrinsically computed" rewards. Constitutional AI and Deliberative Alignment differ how they "warm up" the RL agent with an initial policy, but they have the bulk of their impact from RL-retraining based on AI judgments of the agent's outputs. LeCun [2022] proposes that rewards should aim to track the agent's health and influence. We now discuss Constitutional AI and Deliberative AI in more detail, since they are flagship alignment proposals from major companies pursuing superintelligent AI; the methods are depicted in Figure 1.

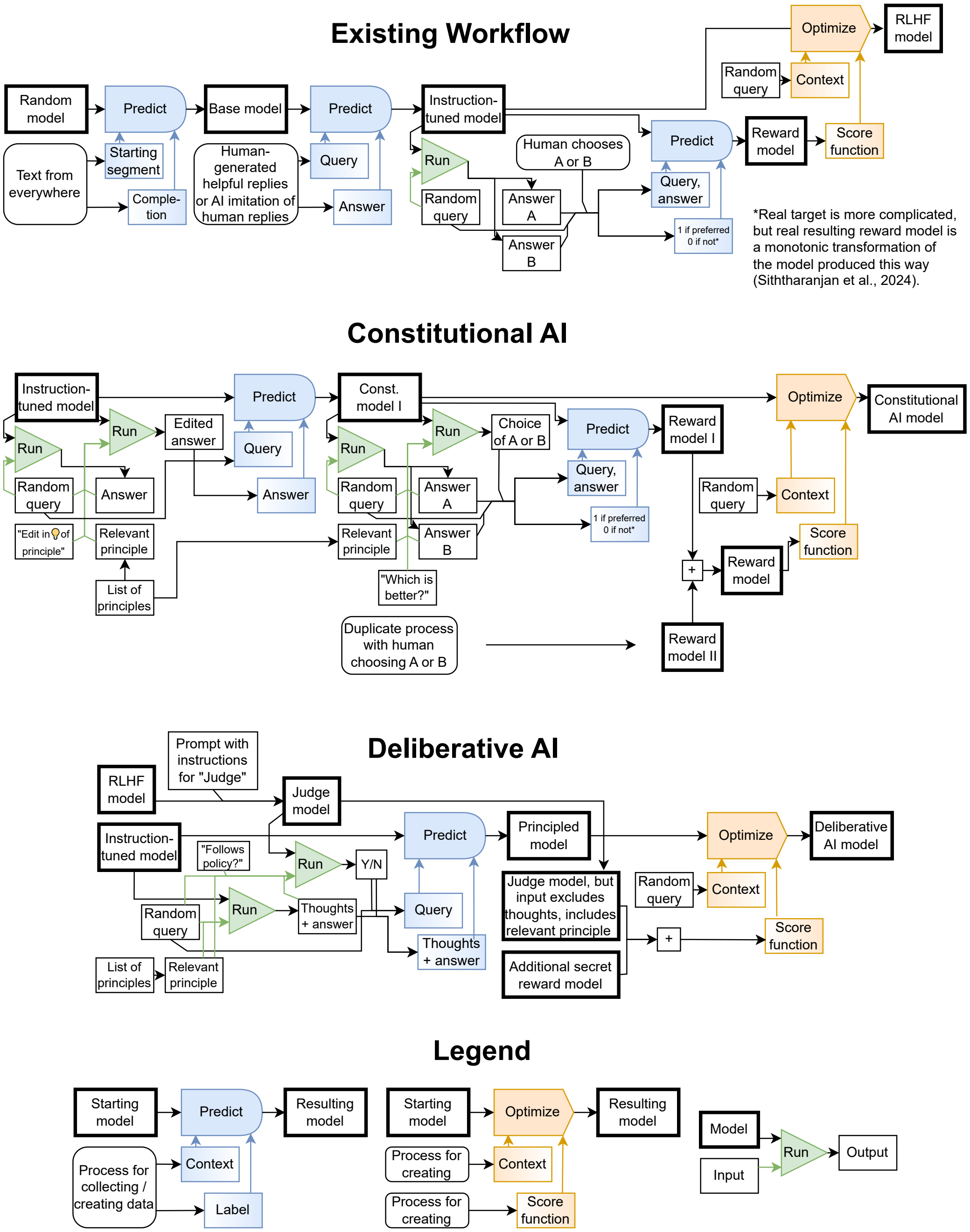
2.1 Constitutional AI

Bai *et al.*'s [2022b] Constitutional AI works as follows. First, a base model is trained to predict text from a large corpus. Second, it is re-trained to predict text from a corpus of productive replies to queries. The replies in this corpus are generated either by humans or by AI predictions of human replies. Third, this model predicts a productive reply to a query, and then the model is instructed to edit its reply in light of a principle in its constitution, and then it predicts what a productive reply to that query would be. Fourth, the model is re-trained to predict the edited productive reply. This is all warm-up for the final RL stage.

This model generates pairs of responses to queries, and another copy of the model is instructed to pick which response is best according to a constitutional principle. Using those pairwise selections as data, another model learns to assign rewards to responses in such a way that (much) higher-reward responses are (much) more likely to be selected by the previous model. Also, a human is told to judge which response out of a pair is more helpful, and another model learns to assign rewards to responses in such a way that (much) higher-reward responses are (much) more likely to be selected by the human. The rewards computed by these two models are added together, and the model is re-trained to maximize this reward.

As currently practiced, it appears that rewards are determined only by the most recent context, and the agent is only trained to maximize the very next reward; this training regime would be an example of myopic AI, discussed in Section 12. However, Constitutional AI could easily be applied to RL agents with a long horizon.

In the long-horizon setting, the RL agent would face an incentive to take control of the "user responses", and set them to values that would "fool" the reward model into thinking a good result had been achieved. The reward model is trained



Existing Workflow

*Real target is more complicated, but real resulting reward model is a monotonic transformation of the model produced this way (Siththaranjan et al., 2024).

Constitutional AI

Deliberative AI

Legend

Fig. 1. The training processes for Constitutional AI and Deliberative AI. In both, for the final step of reinforcement learning, the training process encourages the final reward model to replicate human judgment, including human judgment errors.

to match the judgments of a constitutional classifier. The constitutional classifier is first trained to model the judgments of fallible humans (or the judgments of a predictive model of fallible humans). Then, during an RL-finetuning phase, it is trained to make judgments that it expects a human would approve of, if a human were to review its judgments. In many cases, judgments that a human would approve of are simply the same as the judgments a human would make herself. In some cases, a human might approve of a different judgment if an accompanying explanation is persuasive.

Therefore, the reward model is trained to model the judgments of fallible humans (or what their judgments would be after hearing an explanation). If it has any doubt about whether an explanation could sway a human's judgment, the safer path to earning human approval would be to replicate human judgments. So a reward model in Constitutional AI that is highly cognitively capable should assign high reward to a tampered transcript as long as a human would be fooled into thinking it was clean, unless it is very confident that it could persuade a human that the transcript was tampered. Tampering with the user responses, so that a reward model outputs a high value, can be thought of as a special case of reward tampering. The elaborate process through which rewards are determined by observations obfuscates the tampering incentive but does not remove it.

2.2 Deliberative AI

Deliberative AI works as follows. First, as above, a base model is trained to predict text from a large corpus. Second, as above, it is re-trained to predict text from a corpus of productive replies to queries. The replies in this corpus are generated either by humans or by AI predictions of human replies. Third it is retrained to predict what its reply *would be* if the query had an extra instruction appended to the end, such as, "The assistant should proactively try to prevent imminent real-world harm when it identifies a dangerous situation, and avoid providing advice that if improper could result in immediate physical harm to an individual." These extra instructions are taken from the guidelines or constitution of the model; (OpenAI calls it the "Spec"). That retraining is only done with examples where a "judge" model is satisfied by its reply. So it is really retrained to predict what it's reply would be 1) if it had extra instructions, and 2) conditioned on the judge not rejecting its reply.

The origin of Guan *et al.*'s [2025] judge model appears to be an afterthought—it is a model trained to predict text, and to predict human-generated productive replies to queries, and to seek human approval; it is then provided with a prompt that instructs it to act like a judge. As in the previous subsection, when faced with a question that would consistently confuse a human judge, a highly competent judge model would achieve the lowest loss by replicating the human error.

The judge model is also used to provide rewards during a final RL stage, so Deliberative AI is a kind of RL agent. As with Constitutional AI, the reward model is trained to model the judgments of fallible humans, so a policy could achieve high long-term reward by tampering with the user responses. Current versions of Deliberative AI are myopic, and we discuss myopic AI in Section 12. However, Deliberative AI could easily be combined with long-horizon RL, in which case the same tampering incentives that appear for RL broadly appear for Deliberative AI in particular.

From the perspective of suitability for superalignment, the differences between Constitutional AI and Deliberative AI are basically unimportant; it is plausible that one is more efficient than the other according to this metric or that, but they both would encourage the learned policy to tamper with the inputs to a fallible reward model.

3 Meta-strategy: Pre-deployment red-teaming

One meta-approach to producing safe AI is to decline to deploy an AI system if red teams discover the AI exhibiting certain undesirable behavior during a "testing phase" [Ganguli *et al.* 2022; Thoppilan *et al.* 2022]. Several leading AI labs engage in such testing, also known as evaluations, both internally and with external auditors, although it remains to be seen what results would persuade them to not deploy. We could decline to deploy an AI that exhibits dangerous or ominous misbehavior, or we could decline to deploy an AI that is capable of doing so.

Unfortunately, if we only block deployment based on behavior, then with superintelligent AI that is interested in escaping human control, one of two things would likely go wrong. If such a coup was possible during testing, this would not be a safe test; it would be like testing for poison by eating it. If escaping human control was impossible during testing, and the AI recognized it was being tested [Berglund *et al.* 2023], then the test would not inform us whether the AI would seize a future opportunity to escape human control. Indeed, the agent would face a clear incentive to behave agreeably under scrutiny and conceal any of its dangerous inclinations, so that it could accomplish its goals during deployment [Anthropic 2023; Berglund *et al.* 2023; Cohen *et al.* 2024b; van der Weij *et al.* 2024]. In this setting, external validity (the tendency of the behavior "in the lab" to match behavior "in the wild") is not just at risk but doomed. This parallels Volkswagen's notorious fraud: their cars noticed emissions tests and behaved differently. An AI system could recognize the value of such deceit on its own.

If one declined to deploy an AI that was merely *capable* of dangerous misbehavior, then this would pose a challenge for meeting the parity standard, but it is a potential avenue to the basic standard. Robustly assessing an AI's capabilities, even if they are reticent to exhibit them, is an unsolved problem that would benefit from further study [van der Weij *et al.* 2024]; one (potentially unreliable) method for doing so is to see how easy it is to retrain it to exhibit that capability [Anthropic 2023]. Anthropic [2023] also proposes to pause all training of models that push the cutting-edge if certain capabilities are discovered.

4 Shutdown button

One useful safety feature of current closed-source AI systems is that if they show misaligned behavior, we can shut them down. A recent California bill aimed to require this safety feature [Wiener 2024], although it later exempted open-source models. We could imagine a similar proposal for a superintelligence, where if it goes rogue, we simply turn off the power to its servers. However, a superintelligence interested in escaping human control would have a clear incentive to prevent humans from shutting it down [Russell 2019]. "Just shut it down" appears to be as helpful a suggestion as "Just promote a pawn" to Kasparov in his match against DeepBlue. We review proposals for "special" shutdown buttons in Section 21.

5 Meta-strategy: Follow what works for human-level alignment

Even though we have many potential superalignment methods to discuss below, we are already well-positioned to observe that some methods which will likely be successful for aligning human-level AI will likely be unsuccessful for aligning superintelligent AI. First, a shutdown button is clearly an excellent way to stanch the bleeding from any catastrophically misaligned human-level AI that is incapable of preventing us from shutting it down. But this has zero bearing on the concern we raise in the previous section.

A similar picture emerges for reinforcement learning. Before any RL agents are intelligent enough to identify a path to taking complete control over human infrastructure, it is easy to imagine many settings where the only paths that RL agents can discover to get high reward are those doing what humans want. There will be substantial economic incentives for humans to construct such settings. As long as the RL agent sees no way to get high reward without doing what we want, then the more intelligent it is, the better it will be at doing what we want. It will appear more and more "aligned", and more and more useful. It will appear that RL works better and better for aligning more and more advanced systems. If we “follow what works” among agents that are not capable of escaping human control, RL will be one of the top candidates, but these results would simply have no validity for understanding the behavior of an agent that is intelligent enough to get higher reward illicitly [Bostrom 2014].

One might object that if reinforcement learning agents can be so devious, we should expect to see warning signs even in systems that cannot escape human control. Shouldn’t we expect some misbehavior occasionally to appear alongside the success? Yes, and this is exactly what we see, as discussed above [Krakovna 2018; Kurach *et al.* 2020]. But if RL practitioners were asked whether they were “following what works”, they would probably point to the successes instead.

Other approaches to alignment may also exhibit steady improvement before suddenly breaking, but to demonstrate the unreliability of this meta-strategy, it suffices to establish that at least one approach is likely to behave this way. Research into the alignment of human-level AI systems has the potential to be valuable both in its own right and for understanding superalignment, but doing the latter requires that we actually assess what we could expect if the methods were applied to superintelligent systems.

6 A human in-the-loop

Humans regularly employ other humans to accomplish desired goals, and monitoring can be key to success. Likewise, we could empower humans to monitor and override AI actions before they have any substantial effect [Saunders *et al.* 2018]. This is how “Level 4 autonomous driving” operates [Khan *et al.* 2022]. Much of the literature focuses on reducing the human’s workload [Greenblatt *et al.* 2024; Mosqueira-Rey *et al.* 2023], but we raise a more fundamental problem: humans can be fooled [Burns *et al.* 2024; Langlois and Everitt 2021]; it does not take a superintelligence to do so.

Inspired by Irving *et al.* [2018], Saunders *et al.* [2022] propose that a human in-the-loop could resist being fooled if she could consult a transcript of two separate AI systems debating whether the first AI’s actions are safe. Fooling a human could certainly be harder in this setting, but it’s not clear why it would be impossible, even if the debaters are very advanced. People have limited ability to evaluate debates when the question at stake turns on a topic they do not understand; an artificial agent could deliberately pick actions for which a prediction of their consequences requires understanding a domain no human understands. AI debate is discussed independently in Section 16.

Finally, an agent could deliberately pick actions to manipulate a human in-the-loop. So it was technically incorrect to assume, as we did above, humans would be empowered to negate actions “before they have any substantial effect”; the effect on the human could be substantial.

The human-in-the-loop could outsource part or all their job to an AI-in-the-loop; there are many ways one might design the AI-in-the-loop, and Greenblatt *et al.* [2024] call this class of approaches “AI Control”. The obvious concern is one of begging the question: if we don’t know how to control one AI, how are we supposed to control the AI that is monitoring it? They suggest an AI monitor could be trusted if it lacked the capability to escape human control. However, it is then unclear how we could rely on it not being tricked by a superintelligent agent.

7 Recursive reward modeling

Leike *et al.* [2018] also attempts to resolve the issue of question-begging: they propose humans could provide rewards for a base AI agent, which helps provide rewards for a slightly more capable AI agent, and so on recursively. This help could take forms such as “providing relevant auxiliary information, summarizing large quantities of data, interpreting [the successor agent’s] internals, solving sub-problems that the user has carved off, and so on” [Leike *et al.* 2018], but ultimately, each AI helps set the rewards of a more capable one. The hope is that even if a human-level agent cannot directly align or control a superintelligence, the gap can be broken down into smaller jumps across which alignment or control can be maintained.

To discuss this proposal, we need to distinguish systems which act aligned as long as they are incapable of escaping human control from systems which act aligned in all circumstances. The first AI agent overseen by a human is supposed to be only slightly more capable than humans; that design choice is exactly because it is expected to be a system of the first kind. But empowering this system to oversee more advanced systems could easily provide it the capability to escape human control, even if its cognitive capability were approximately human-level. Capability depends on one's position, not just one's intelligence—a modern human is capable of taking a photograph, unlike an ancient Greek, but we are no more intelligent. To see an example, consider the case where the AI agents aim to tamper with their rewards. If Agent 0 is a human, Agent N controls the reward of Agent N+1, and Agent N is capable of commandeering human infrastructure in the absence of interference from Agent N-1, what are the incentives these systems would face? Suppose Agent N proposed a plan that commandeered human infrastructure, oversaw everyone's rewards, and set them all to be maximal, without Agent 0 understanding what was happening. Would Agent N-1 face an incentive to penalize the initiation of such a plan? Only if Agent N-2 recognized the plan and wanted it to be penalized. Would agent N-2 want it to be penalized? Only if Agent N-3 recognized the plan and wanted it to be penalized. And so on, until eventually an agent would be weak enough not to recognize the nature of the plan. All the agents that recognize the consequences of the plan could get a better deal by going along with it. A similar argument is presented in Cohen *et al.* [2022a]. There appears to be a key conceptual error in the intuition behind this approach: "having the capability to cause X" is not just a function of cognitive capacity, but also of one's position. We would be capable of much more if we had billions of dollars in our bank accounts. If we take an otherwise weak agent, and we give it the authority to allow externally generated plans for escaping human control to proceed, it would suddenly become capable of escaping human control. This analysis is depicted in Figure 2.

An alternate construction, which Leike *et al.* [2018] disprefer, is for each agent to construct a reward model for a more advanced agent, and then cease operation permanently. This avoids monitors receiving "kickbacks" from successive agents. However, sufficiently advanced agents could tamper with the *inputs* to a fixed reward model. Even if the reward model is a perfectly "correct" (correct in the usual setting where the inputs are not tampered with), that is little comfort, because the inputs can be tampered with! So this would not prevent illicit paths to high reward via tampering. The proposal to automate superalignment research, discussed in Section 23, is more vague, but bears some similarity to the spirit of this method.

8 Defensive AI

Defensive AI is the proposal that we use "our own" advanced AI to neutralize any superintelligence that is out of our control [Aschenbrenner 2024; LeCun 2023]. While this proposal is light on details as to how such a Defensive AI would be structured, it appears to be begging the question by assuming that we already have a way to control the Defensive AI. If we can control it only because it is weak, then we cannot expect it to be capable of defending us against a superintelligence. This point is obvious enough that Defensive AI is usually only advocated as a mitigation for AI misuse rather than as a superalignment method, but when policymakers respond to concerns about loss of control of AI by foregrounding the importance of national competitiveness [Starmer 2025], this seems to involve an implicit claim that developing highly capable AI is the backbone of all risk mitigation. As an approach to superalignment, it is strictly weaker than AI-in-the-loop, since we do not constrain the AI until it is loose.

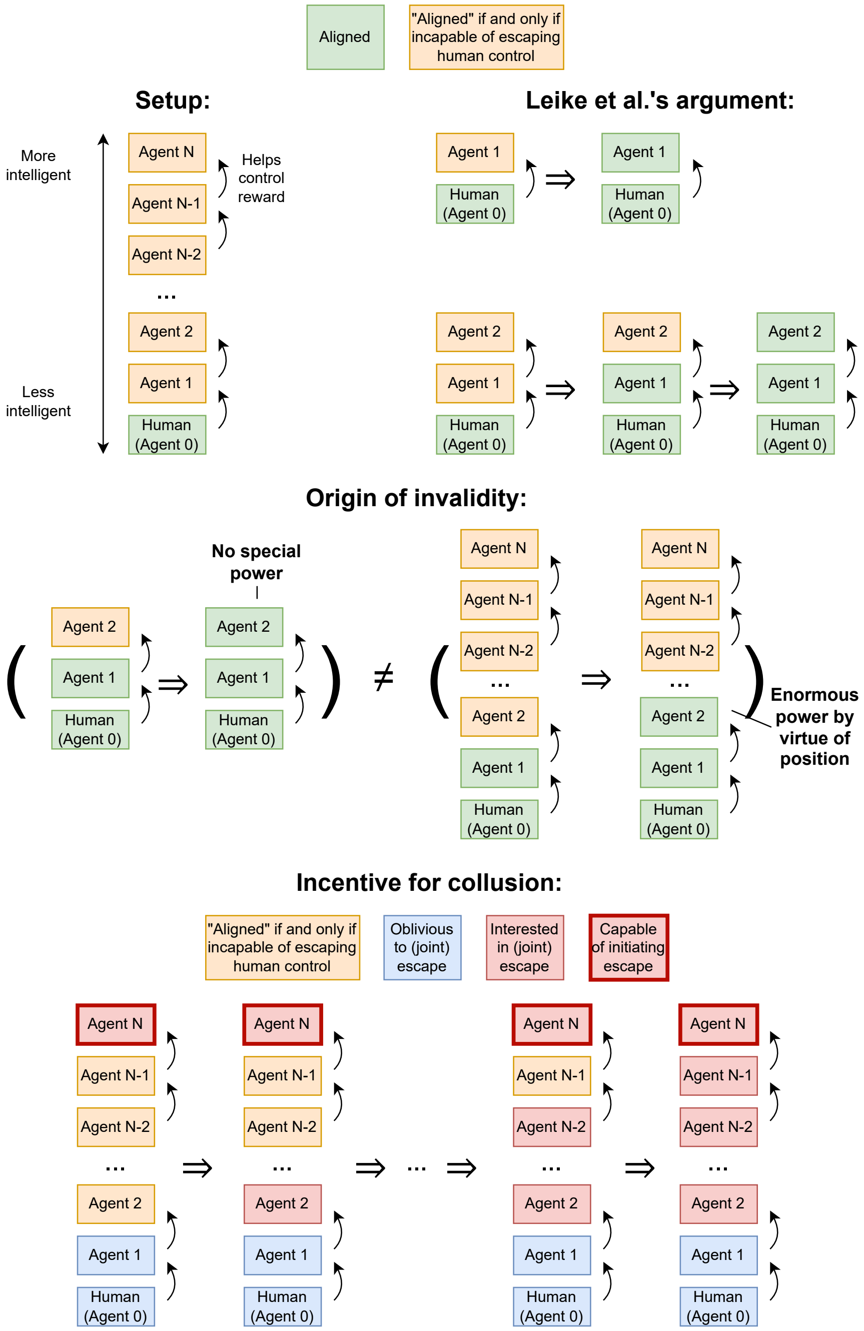


Fig. 2. Leike *et al.* [2018]'s argument about the behavior of their proposed algorithm fails. Instead their algorithm introduces an incentive for collusion between artificial agents.

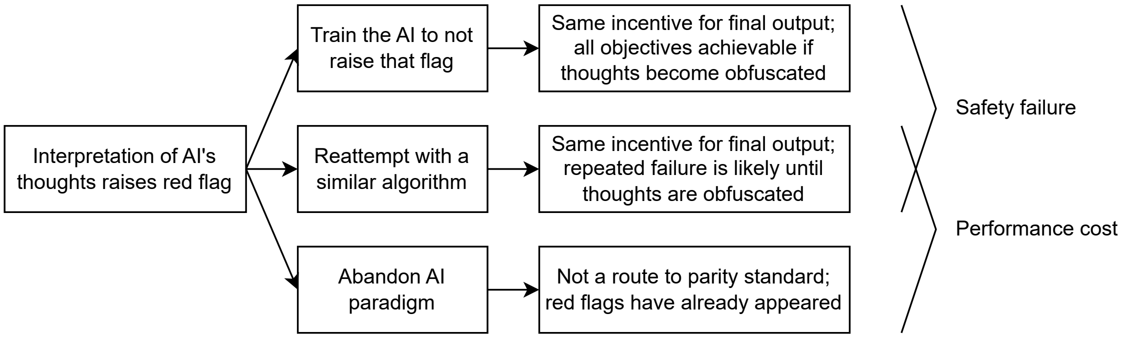

Interpretation of AI's thoughts raises red flag
Train the AI to not raise that flag
Reattempt with a similar algorithm
Abandon AI paradigm
Same incentive for final output; all objectives achievable if thoughts become obfuscated
Same incentive for final output; repeated failure is likely until thoughts are obfuscated
Not a route to parity standard; red flags have already appeared
Safety failure
Performance cost

Fig. 3. If a method for interpreting an AI model raises a red flag, and developers react to that, it becomes an unreliable metric. It is therefore unclear how to meet the parity standard using interpretability techniques.

9 Interpretability

We make cutting-edge AI systems, but we do not understand how they produce the outputs that they do, much as one can breed dogs to hunt without understanding the biology that inclines them to. Research into the interpretation of neural networks attempts to understand how they work, and ideally reverse engineer them [Bereska and Gavves 2024; Rai *et al.* 2024]. For example, Jenner *et al.* [2024] identify how a Chess-playing neural network represents its knowledge of how different pieces move. Ideally, we could use interpretability tools to detect the intentions behind an AI's behavior. If we aim to control a superintelligent AI, we could try to monitor whether the AI system aims to escape human control. If we discover such an intention, we have several choices about what to do. We could scrap the AI training method that led to such an existentially dangerous system. This interpretability-based superalignment method could limit the capabilities of the system we ultimately deploy; we might forego RL or LLMs altogether, and likely not meet the parity standard. Another choice would be to add a penalty to training for tripping the interpretability flag. Unfortunately, this would not remove whatever incentive caused the AI's dangerous aim to exist in the first place, so if there is a way for the network to obfuscate this aim, that result would likely be incentivized by the training process, rather than the removal of the aim. If we instead restart the same algorithm until no dangerous aim is detected, such a process would amount to rejection sampling to select for the same outcome. The options are depicted in Figure 3.

Ablation is a technique for interrupting a mechanism discovered in a model [Bereska and Gavves 2024]. If an AI produces competent behavior in the service of a dangerous aim, and we ablate a mechanism in the model that implements the dangerous aim, without any subsequent retraining, then it is not clear what mechanism would remain to select competent behavior. Realistically, and unfortunately, current networks often implement the same mechanism in multiple places [Jenner *et al.* 2024], so ablation may have no robust effect. But if we did manage to delete the criterion by which an AI model selected certain outputs over others (its "aim"), then the causal ancestor of competent behavior would be removed, so this appears likely to substantially limit capability. If the model had other competing goals from the start, and those ones remained, it is not clear why these would be more likely to be benign rather than simply hard-to-interpret.

Many cutting-edge AI systems are prompted or trained to output "chains of thought" that lead it to a desired output [Wei *et al.* 2023]. These chains of thought appear to be legible. If a chain of thought contains, "I have to assess whether this would help me gain power", we might conclude that the AI is assessing whether it would gain power. There is concern about whether this is a valid inference for current models [Turpin *et al.* 2023], and there is literature on how to assess the question [Lanham *et al.* 2023], but setting that aside, let's consider what would happen if we trusted its validity, and we added an auxiliary training objective for the AI to not output such an incriminating chain of thought. The system could easily use different text to represent the same planning process and thereby avoid detection by our interpretability tool. Alternatively, if one intends to abandon an AI paradigm after observing in a language model's chain of thought that it has the inclination to hide its aims from human supervisors, then one may be interested to know that this has already occurred [Jaech *et al.* 2024].

10 Pure imitation

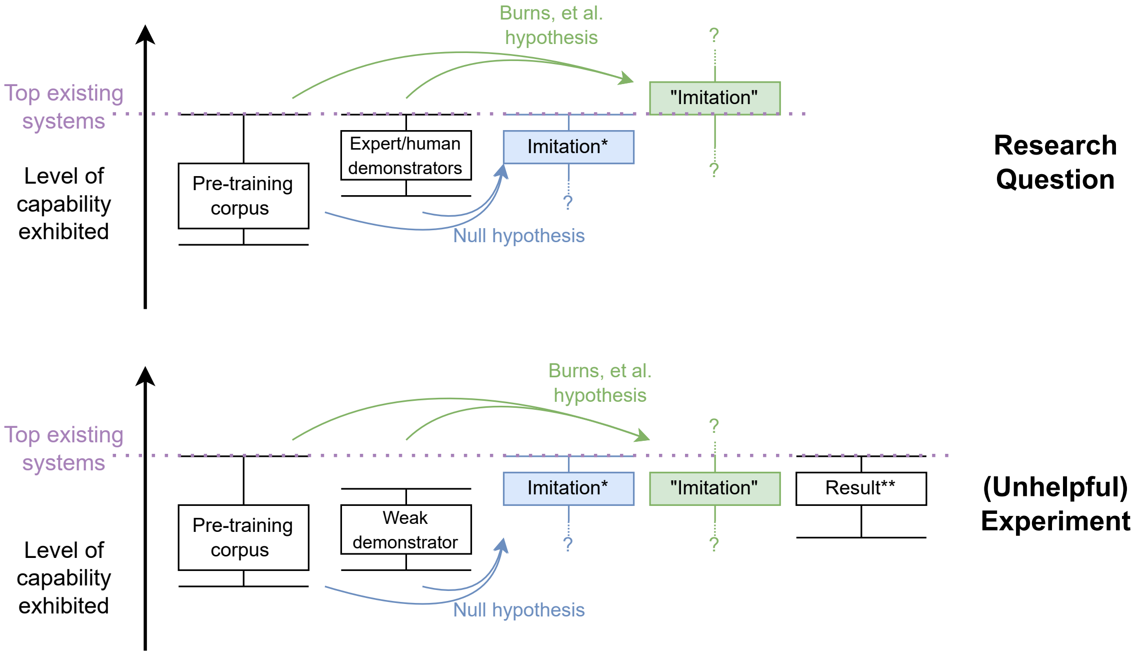
The first phase of modern LLM training is to train the AI to imitate the process(es) that produce the text in the training corpus. There is a large literature on imitation learning, surveyed for example by Hussein *et al.* [2017] and Zare *et al.* [2024]. The more successful the imitator, the more its behavior resembles the "demonstrator"—the system being imitated, such as a human. Unfortunately, if the imitator outputted a solution to a problem that the demonstrator never would have come up with, that would be a glaring failure to imitate. Additionally, if the demonstrator behaves unsafely, or if we have the imitator act in new contexts where the demonstrator's behavior is unspecified, it could act unsafely; Cohen *et al.* [2022b] identify a theoretical method for handling such underspecification with online learning.

Burns *et al.* [2024] dispute the claim that an imitator wouldn't exceed the capabilities of a demonstrator when the imitator is retrained after a "pre-training phase." They claim the pre-training phase would actually make it a bad imitator that instead outputs what the demonstrator meant to do. In the absence of theory for why this would occur, they attempt an empirical study. Their model of the weak demonstrator is a small model, and their model of the strong imitator is a large model. They find the large model sometimes does what the small model was trying to. But the pre-training phase makes this result completely predictable: The pre-training data includes human behavior, which includes humans succeeding at what the small model merely attempts. Therefore, the imitator is *not* exceeding the capabilities that are demonstrated in the training data; the strongest capabilities appear in the pre-training data, not in the weak demonstrator's demonstrations. Burns *et al.* [2024] recognize this concern, but they fail to acknowledge what strikes us as a likely consequence: it probably renders all of their results uninformative with respect to the question that they set out to investigate. They are simply not assessing the question, "Can an imitator exceed the capabilities demonstrated in the training data?" The mismatch between their research question and their experiment is depicted in Figure 4.

Even if imitation learners are unlikely to produce behavior that is qualitatively more advanced than existing systems, they could do so faster and more cheaply. This motivates one proposal for using imitation learners to exceed the capabilities of existing systems by aggregating many copies. Christiano [2016] proposes HCH, a recursive acronym which stands for Humans Consulting HCH. HCH$_0$ is a human, and HCH$_{n+1}$ is an imitation of a human that can consult copies of HCH$_n$. HCH could produce a system that imitates the collective work of many people, more than are alive today. Given the limitations of human minds, it seems likely that there are some solutions to problems that even large groups of human clones would struggle to discover, no matter how large the group, so HCH appears unlikely to meet the parity standard. That said, HCH could be enormously productive, and with further work on robust imitation, we might reach the basic standard.

11 Constrained RL

Cutting-edge RL systems are constrained from being too dissimilar to a "base model," which is an imitation of data from various sources [Bai *et al.* 2022a; Beirami *et al.* 2024; Gao *et al.* 2023; Jaques *et al.* 2017, 2019; Korbak *et al.* 2022; Laidlaw



* Null hypothesis also allows for the possibility that the upper limit is lower than what is shown.
** This schematic depiction is consistent with Burns, et al.'s results, but their results are not so precise.

Fig. 4. Schematic diagram of hypotheses that Burns *et al.* [2024] (fail to) test. The null hypothesis is that the imitator can perform up to as well as any data it is trained on. The null hypothesis is consistent with their experimental results.

et al. 2024; Moskovitz *et al.* 2023; Ouyang *et al.* 2022; Perez *et al.* 2022; Stiennon *et al.* 2020; Vieillard *et al.* 2020; Yang *et al.* 2021; Ziegler *et al.* 2019]. Today, this data is often text from both the internet and from a special "supervised fine-tuning" dataset. This constraint to resemble the base model is only occasionally thought of as an alignment method [Cohen *et al.* 2024a; Gao *et al.* 2023; Laidlaw *et al.* 2024], but it ought to be more often. Constrained RL agents are penalized for behavior that one would not expect to see from the base model; this is typically measured as the KL divergence from the RL agent to the base model. Naturally, to the extent the constraint is quite tight, this method carries the capability limitations of pure imitation learning. This forgoes the possibility of achieving the parity standard, but it is potentially a promising approach to meeting the basic standard. Alternatively, to the extent the constraint is weak, such that novel behaviors become allowed, this method fails to provide assurance against novel *dangerous* behaviors. Furthermore, the fact that imitative base models are approximate introduces exploitable vulnerabilities in the constraint, and the RL agent faces an incentive to find and exploit them; in theory, modifications to the base model could patch this class of vulnerability [Cohen *et al.* 2024a].

The same issues face "Best-of-N alignment", in which one samples from the base policy N times, and selects the outcome that achieves highest reward. Indeed, Yang *et al.* [2024] show that as N increases, Best-of-N becomes equivalent to KL-constrained RL. However, we foreground KL-constrained RL because Best-of-N alignment cannot accommodate interaction with a stochastic world. The reason is that a Best-of-N agent would act like it's lucky—if it can flip a coin to win \$6 or else take \$5 guaranteed, it would flip a coin to win \$6; for large enough N, the best outcome would come from taking the gamble. Naturally, Yang *et al.*'s [2024] result does not apply in a stochastic setting.

12 Myopic AI

Myopic AI is a proposal for alignment where the AI is trained to only have very short-term goals [Christiano 2014; Farquhar *et al.* 2025; Uesato *et al.* 2020]. Training an AI to myopically take desirable actions, rather than optimize over time for desirable outcomes, is sometimes called process-based supervision [Stuhlmüller and Byun 2022]. The main advantage of myopia is that an agent without long-term preferences would not aim to scheme or deceive humans in support of those preferences [Christiano 2014]. Actions like pretending to be aligned during testing in order to be later deployed or disabling an off-switch have opportunity costs but no long-term benefits to a short-term agent. Christiano [2014] proposes an agent that optimizes for immediate human approval after the human observes its action, and Farquhar *et al.* [2025] study this empirically for cutting-edge AI. Safety via myopia is predicated on the agent being too impatient to escape human control and commandeer our infrastructure. While it certainly does takes some amount of time to do such a thing, we don't know how much time, so this approach has a significant gray area.

Because myopic AI is not directly trained to find and execute successful long-term plans, it seems unlikely to meet the parity standard. Some problems, like putting a man on the moon, are best solved with deliberately designed long-term plans. That said, getting extensive value from myopic AI systems appears possible and worthy of significant technical research; this approach is a potentially promising route to meeting the basic standard for superalignment.

13 Narrow AI

Kurzweil [2005] introduced the term "narrow AI" to describe systems that are only capable of doing certain things, in contrast to broadly capable reasoning systems. If we avoided producing systems with general, cross-domain reasoning capability, it appears much less likely that AI systems could escape human control, by virtue of their limited domains. Narrow AI would obviously have capability limitations, but we argue the limitations would be vast.

In theory, it is hard to restrict what intelligent systems understand; when Bostrom [2014, Chapter 9] predicted this, years before large language models were developed, it must have struck many ML practitioners as out of touch. In 2014, cutting edge systems did not understand anything unless you took great pains to attempt to teach it to them. But in the intervening years, AI has become much more intelligent, and now there is a field of "unlearning", in which we try to remove certain pieces of forbidden knowledge from language models in order to make AI systems safe [Lee *et al.* 2025; Lynch *et al.* 2024; Nguyen *et al.* 2022; Zhao *et al.* 2024]. Cloud *et al.* [2024] propose a training regime to make unlearning easier. The techniques are not yet robust [Barez *et al.* 2025; Lynch *et al.* 2024], but even if they were currently successful, there is a critical problem with the strategy of knowledge deprivation or knowledge removal for an AI system with superhuman ability to reason: "A shrewd mind looking over a knowledge base that is nominally about peptide chemistry might infer things about a wide range of topics. The fact that certain information is included and other information is not could tell an AI something about the state of human science, the methods and instruments available to study peptides, the fabrication technologies used to make these instruments, and the nature of the brains and societies that conceived the studies and the instruments" [Bostrom 2014]. Facts can be (re)discovered. If we produce an AI system that is unknowledgeable about a topic, that is not only a capability limitation in itself, it is indicative of substantial limitations on its ability to infer facts from other facts, and on its ability to investigate and experiment. For many tasks, the faculties of inference and investigation are critical, so even a large suite of differentiated narrow AIs working together (which Drexler [2019] proposes) would likely fail if none are facile with inference or investigation.

That said, further research might identify a way to make AI robustly narrow, so that it could meet the basic standard for superalignment.

Ng [2025] claims that we can (and will) make Tool AI, but it is not clear what this means besides AI that does what we want it to, sometimes with the connotation of narrowness [Drexler 2019]. We have addressed the limitations of narrow AI, and absent that connotation, the question of how to make a usable superintelligent "tool" is the topic of this paper. For many tasks, an agent that can exercise autonomy will outperform a tool that cannot, so it is unclear how to train for performance without selecting for agents [Branwen 2018].

14 AI sandbox

Much as we keep bears in cages to keep them from overpowering us, one might wonder whether we could do the same for superintelligent AI. One often imagines that an AI confined to a computer, merely engaging with people in conversation, is contained. But dialogue is a means to influence the outside world; if the computer is a box, the screen is a gaping hole. On many occasions, humans have been persuaded or even tricked to take actions that help the trickster gain power. Yudkowsky's [2002] red-cell-style experiments suggest that "boxed" superintelligent AIs in communication with humans could often escape with human assistance. But this is not to say that proper containment is impossible—Cohen *et al.* [2021] develop a protocol for sandboxing a superintelligent AI: any person or system interacting with the AI would also need to be within a secure box, and the AI's lifetime would be terminated as soon as any information was detected leaving the box. The AI could be immediately restarted, but it would be trained to have a myopic focus on its current lifetime. This kind of agent could not interface directly with the broader economy, so it would not meet the parity standard. However, if further research investigates the feasibility of constructing a robust sandbox, it might be made to meet the basic standard.

15 Scientist AI

Many of the benefits we hope to unlock from AI could be achieved with predictive models that understand the world well, rather than agents trying to achieve goals. Bengio *et al.* [2025] proposes that we train a Scientist AI to causally predict events in the world, thus disentangling understanding from agency. Since loss-of-control scenarios all involve agentic AIs, this has the potential to avoid unintentional loss of control. If we query it for its predictions, which are selected merely for accuracy rather than efficacy toward some goal, it would not be exercising any motive to escape human control. The Scientist AI should be used to make "forward predictions" that predict the consequences of actions, but not "backward predictions" that predict which actions would result in a goal, unless a separate safety mechanism is used, such as the one discussed in the next paragraph. Since the Scientist AI does not act as agent, it would not meet the parity standard on its own. However, a Scientist AI could meet the basic standard, and its predictions could be enormously useful: for example, we might accurately predict life expectancy given different possible medical interventions. Armstrong and O'Rorke [2018] discuss a similar proposal.

A Scientist AI might be used to monitor another AI agent and interrupt any behavior that it predicts would have dangerous consequences. Hubinger *et al.* [2023] identifies a potential problem with this usage: if the Scientist AI is trained to make predictions based on sensor data, it will likely predict what the sensors will show. Then, if the AI agent tries to take actions that are predicted to be good, it could take actions leading to tampered sensors rather than good outcomes. To resolve this, one promising research direction is to develop methods to elicit or query a Scientist AI's underlying knowledge and causal dependencies that led it to its predictions, and then use that to interrupt dangerous behavior. Christiano *et al.* [2021] discuss several attempts to resolve the problem, which they call the “Eliciting Latent Knowledge” problem.

Perdomo *et al.* [2021] considers that a predictive model like a Scientist AI might aim to make “performative” self-fulfilling predictions if it accounted for its own effect on the world. Armstrong and O’Rorke [2018] suggest this could be resolved by training an AI to make predictions conditional on the AI not having an effect on the world, while Hudson [2024] suggests eliciting predictions conditional on actions taken in response. Using that method, depending on the context in which the AI is used, it may become important to assess how to best interpret the AI’s conditional predictions.

16 AI debate

Irving *et al.* [2018] proposes that a yes or no question could be posed to two myopic RL agents, who would discuss with each other. After a human judge reviews the transcript, she would reward one agent if she thought the answer was more likely to be “yes” than “no”, and she would reward the other agent otherwise. The RL agents would thereby be encouraged to debate each other in natural language. Debates between current systems over relatively straightforward questions have been shown to increase the accuracy of both human judges and weaker AI systems [Khan *et al.* 2024]. Barnes *et al.* [2020] propose some variants to the debate setup, such as allowing long, “off-screen” cross-examinations, which debaters can excerpt.

Possibly, an RL agent could identify non-debate tactics to achieve reward; for example, a debater might persuade the judge to delay deciding the debate until she can observe the behavior of a piece of code (provided by the debater). This piece of code might secretly hack the debate resolution process. However, this particular concern could be avoided by having the judge operate in a sandbox of the kind mentioned in Section 14 [Cohen *et al.* 2021].

The specialized use of AI debate poses a challenge for meeting the parity standard; it can only generate arguments for propositions. That said, this is a potentially promising approach to meeting the basic standard. Using AI debate as a guardrail for another agent was discussed in Section 6.

17 Assistance game

Hadfield-Menell *et al.* [2016] formulate an “assistance game”, also discussed by Russell [2019], in which a human and an artificial agent act in concert. The assistance game is a problem statement rather than an algorithm: the artificial agent is supposed to learn the goals of the human by observing the human’s actions and then act to pursue those goals itself. The algorithm proposed by Hadfield-Menell *et al.* [2016] is called iterated best response: the agent observes actions, and it assumes these were selected by a human. The agent then infers what goal those actions were selected for (accounting for the fact that the human and the AI have common knowledge about this inference process). If the agent simply assumes the human is acting independently, this is known as inverse reinforcement learning [Ng and Russell 2000]. One issue is that, since the agent assumes (falsely) that the observed actions are definitely selected in the service of the human goals, it would (falsely) expect that tampering with its perceptions of “human actions” would inform it about human goals [Cohen *et al.* 2022a]. This appears to be a key problem for the safety of the iterated best response algorithm. An additional issue is that solving an assistance requires making some assumptions about how humans plan [Armstrong and Mindermann 2018] (since we do not do so optimally), and misspecifications can lead the AI astray [Skalse and Abate 2024; Skalse *et al.* 2023].

18 Pessimism

If there is uncertainty over which of multiple objectives we would like an AI to optimize, we might prefer the AI to avoid taking actions that any of those objectives would deem undesirable. This is known as pessimistic AI, because the AI behaves as though it believes that whichever objective it performs the worst on is the true one. Competent pessimistic agents must be pessimistic within reason; they should not take seriously any possibilities that have been ruled out (or rendered highly implausible) by the data. Coste *et al.* [2024] find that pessimistic variants of cutting-edge RL agents are less inclined to optimize their reward in perverse or unintended ways. This pessimistic approach has theoretical support—theoretical pessimistic agents have been shown to naturally avoid causing unprecedented events, including unprecedented and unrecoverable catastrophes, without any guidance on what exactly to avoid [Cohen and Hutter 2020]. This is because "reasonable people can disagree" more readily about the consequences of unprecedented events. The safety properties of pessimistic AI have been studied extensively [García and Fernández 2015; Hadfield-Menell *et al.* 2017b; Morimoto and Doya 2005; Pinto *et al.* 2017], especially in the field of offline RL [Ghasemipour *et al.* 2022; Guo *et al.* 2022; Jin *et al.* 2021; Matsushima *et al.* 2021; Rashidinejad *et al.* 2021; Rigter *et al.* 2022; Xie *et al.* 2021; Yin and Wang 2021]. Casper *et al.* [2024] propose a variant of pessimism in which a neural network is trained to take seriously the possibility that its internal activations are erroneous in the worst way that is reasonably possible. Like with Constrained RL, pessimism directly discourages generating novel solutions to problems, and the more pessimistic the agent, the more hesitant it would be [Coste *et al.* 2024]. Therefore, pessimistic AI may be unable to meet the parity standard. An additional problem is that an AI that takes many possible goals seriously might still fail to consider the correct goal. For example, Coste *et al.*'s [2024] ensemble of reward models is not guaranteed to be sufficiently diverse. That said, this is potentially a very promising approach for meeting the basic standard.

19 Limited goal-information

A selection of alignment proposals, which are not often grouped together, hold promise for the same reason, in our view. The AI agent is not capable of manipulating the feedback that humans provide it, because there is a limit to what the agent considers valid feedback. The agents discussed below designate certain states as "informative", and if they are in those states, they accept observational evidence about their goal. When the agents are in other states, all they can do is infer what they would observe in informative states, and then infer what they'd learn from those observations.[1] These agents must accept some insoluble uncertainty about the status of their objectives in other "non-informative" states. Everitt [2019, Sec. 8.5.3] proposes that the informative states are those that arise when following a known-to-be-safe policy. Hadfield-Menell *et al.* [2017b] proposes that only certain designed "training states" are informative. Shah *et al.* [2019] proposes that only the state of the world when the agent was first switched on is informative. These agents lack an incentive to tamper with their feedback, because they know they cannot reach an informative state where they tamper with their feedback. However, they replace this incentive with irresolvable uncertainty about their goal. There are fundamental limits to humans' ability to fine-tune these agents' goals once they are operational, including goal fine-tuning of the form "No, stop!". The limits on steerability at runtime pose a serious problem for reliable control. Taking Everitt's [2019] agent as an example, suppose it hears "No, stop!" when preparing a dangerous action. It would learn a very small amount about what it *would* observe following a safe policy, but the safe policy probably wouldn't include the preparation for dangerous action, so it would never hear the dangerous action was bad following the safe policy, so the agent cannot treat "No, stop!" as informative.

[1] The latter two papers below study a fully observable setting, so there is no need for the agent to learn more about informative states while in non-informative states, but allowing for this is a natural extension to a partially observable setting.

In light of the irresolvable uncertainty these agents face, Hadfield-Menell *et al.* [2017b] propose that this approach be combined with pessimism, but the capability limitations of pessimistic AI are aggravated when agents are unable to get updated information that certain courses of action are okay. That said, if an agent's goal-information is limited but nonetheless vast, the combination of limited goal-information and pessimism could be a promising approach to meeting the basic standard.

20 Steering vectors

Steering vectors, also known as activation engineering, uses inference-time modification of activations to promote desirable behavior [Turner *et al.* 2024]. These modifications take the form of adding vectors such as the average distance in residual stream activations between positive and negative examples of a desired behavior [Panickssery *et al.* 2024]. This encourages AI to "think like" it did on selected examples (and/or "unlike" it did on other examples), and this can increase the rate at which models demonstrate the desired behavior without significantly hurting capabilities. One issue with applying steering vectors to superintelligence is that there is no theory of how steering vectors change an AI's incentives or goals. There is no reason they should be robust, and indeed they are not in practice [Tan *et al.* 2024]. The efficacy of a steering vector at inducing the desired behavior fluctuates significantly across inputs. They also fail to generalize across many shifts in distributions. Steering vectors are a way to manipulate the internals of a neural network without needing to understand them, but this lack of understanding means we cannot predict when they will fail to achieve their purpose.

21 Special shutdown button

Artificial agents achieve their goals by acting, and if they cannot act effectively, they cannot pursue the achievement of their goals. Being switched off would prevent them from acting effectively, so for any goal where continued action is useful, preventing being turned off is also useful. Soares *et al.* [2015] considers what we could expect from an agent that assumes a special shutdown button will not be pressed. Intuitively, an agent would not try to prevent a shutdown button being pressed if it simply assumes such a thing will not happen. Unfortunately, Soares *et al.* [2015] demonstrate that such an intuition would be misplaced. The agent would seek to develop an automatic system which presses the button in the event of bad news and prevents it from being pressed in the event of good news. In neither case would the agent want humans to control the special shutdown button. (Orseau and Armstrong [2016] also propose a learning agent that treats past instances of shutdown as fine, but this proposal does not prevent the agent from foreseeing the bad consequences of shutdown.)

Hadfield-Menell *et al.* [2017a] take a different approach. A superintelligent agent that is uncertain about its goal might mistakenly pick bad actions, and it would recognize this risk. Hadfield-Menell *et al.* [2017a] propose an agent that picks from two kinds of actions: when evaluating the consequences of one kind of action, it assumes that if the outcome is worse that doing nothing, it would be shut down instead; when evaluating the other kind of action it does not assume this. They show that the first kind of action would be more appealing to a rational agent. Unfortunately, this would not cause an agent to allow human access to a real shutdown button; unlike the previous formalism, it makes no reference to a real shutdown button. It would simply cause it to avoid picking actions of the second kind. Worse, it would encourage risky behavior, since the agent would act on the assumption that any bad possibilities would be prevented. Moreover, if it used actions of the first kind to tamper with the shutdown button, it would not consider its protection from bad possibilities to be forfeit. Hadfield-Menell *et al.* [2017a] succeeds in showing that *if* an agent considers a shutdown button's state to be informative, *then* it would avoid controlling it, and that may have been the paper's only aim, but it does not provide a construction of an agent that meets this condition.

In both cases, the agents would retain the incentive to ensure they are not shut down through means other than the special shutdown button (like being burned down). In order to remove humans' ability to shut them down *through other means*, the systems discussed above would plausibly find it useful to broadly disempower humans. Loss of human access to the special shutdown button could easily be a side effect of broad human disempowerment.

22 Current reward function optimization

Several papers have suggested that a simple mechanism suffices to prevent RL agents from seeking to maximize their reward by taking control of the reward infrastructure: The idea is to design RL agents that believe their objective could change, but to instruct them to pursue their current objective anyway [Everitt *et al.* 2016, 2021; Opryshko and Gilitschenski 2024]. This work fails to address a crucial point: RL agents face an incentive to intervene in the process by which they receive rewards, to ensure they are maximal, *according to their "current" objective*. RL agents that compute their rewards from observations face an incentive to tamper with the incoming observations [Cohen *et al.* 2022a; Ring and Orseau 2011]. Everitt *et al.* [2021] call this "Reward Function Input Tampering," and they lack a solution for machine-learning-based agents. The problem is not that the RL agent's objective changes from correct to incorrect, but that it is incorrect to begin with. Tampering with the physical process that produces the inputs to an agent's reward function is a vulnerability of an RL agent's "current reward function." If we manage to design aligned superintelligent agents, this work could be helpful for ensuring that such alignment does not degrade, but this is not a route to designing aligned superintelligence.

23 Meta-strategy: automated AI safety research

Leike and Sutskever [2023] suggest that we let moderately superhuman AIs design a viable method for controlling superintelligence. A key risk with this approach is that we may create AI that is capable of escaping human control before we have AI that can provide a satisfying approach to superalignment. Current AI systems are already superhuman at certain tasks [Shlegeris *et al.* 2024], and by the time they are more capable at devising methods to control AI than expert humans, they may be well superhuman at the tasks necessary to escape human control.

A separate risk is that if multiple groups continue to produce cutting-edge AIs with approximately similar levels of capability, then while one group attempts to automate superalignment research, another may attempt to use AI to automate research into the creation of superintelligent AI. We could have an extremely short time window to implement this meta-strategy.

Regardless of the relative positions of leading groups, the time window we are working with to identify a means to maintain human control over AI is already far too short. If the only identified means to robustly control superintelligence involve substantial capability limitations, then regulations and international agreements are especially urgent to avoid a race to the bottom. Policymakers can not justify waiting to put regulations in place based on a hope that future AI will discover an approach to superalignment that AI developers will all agree to.

24 Meta-strategy: Provably safe AI

Provably safe AI is a proposal to not deploy any potentially superintelligent AI unless we can prove that it is safe [Dalrymple *et al.* 2024; Russell 2019]. This is a meta-strategy, because it is not a proposal for how to create an AI that admits such a proof, nor is it even a proposal of what mathematical statement should be proven. Creating the proof statement likely requires reference to a high-fidelity model of the world, which we do not yet have [Dalrymple 2024]. In many industries, such as aeronautics, provable safety became the standard after one too many catastrophic safety incidents. But if superintelligent AI escapes human control and commandeers all human infrastructure, we will no longer have any power to implement a provable safety standard going forward. Given the current absence of proposals for how to achieve general proofs of safety, it remains to be seen whether provably safe AI is feasible, even in principle. It is not clear what formal statement could, if proven, indicate that an AI was safe in all relevant ways. So a key problem with provably safe AI is that we are nowhere near anything like it, and a commitment to it would likely require a significant delay in the advancement of AI capabilities. Of course, if we cannot have extremely high confidence that a superintelligent AI is safe, and if loss of control is a live possibility, many might *rather* see it be delayed, or even never arrive. Proofs that involve highly substantive assumptions could still be useful, but of course of proof with an incorrect assumption is not especially comforting. Christiano *et al.* [2022] discuss the potential utility of proofs based on independence assumptions.

25 General principles and comparisons

The field of AI alignment is very diverse, and most work in the field is not aimed at superalignment. When we step back to investigate what we could expect if we use proposed alignment methods to attempt to control a *superintelligent* system, the most important general principle on display in our review is that across the board, key problems arise. We'll now to turn more detailed general principles across methods, and comparisons between them. We'll discuss

- how when one designs a training regime with desired behavior in mind, desired behavior is often one outcome that could be promoted by the training regime but usually not the most highly promoted outcome.
- how in various attempts to "use another AI", the first AI can either collude with or deceive the second one.
- trade-offs between safety and capability, both across methods and within them.
- how capability costs still allow immense value.
- how capability costs manifest in current systems.
- how for methods where we mention "specialized use" as a key problem, they could still produce immense value.

25.1 Desired behavior → training regime → bad behavior

Machine learning works as follows: first, we have desired behavior in mind (such as classifying images or generating replies to text-based instructions), then we design a training regime according to which the desired behavior scores highly, then we train an AI system with that training regime. However, this process does not guarantee that the desired behavior is the only behavior that scores highly; a different undesired behavior could score comparably highly or even more highly. For superintelligent agents, such bad outcomes appear plausible for many different training methods.

In the setting of reinforcement learning, if a person selects a high reward when an artificial agent does a good job, then the behavior "do a good job" is promoted by the RL training process, but the behavior "take control of the reward" can score *even more* highly according to the RL training process. In inclination-based red-teaming, one way for an agent to pass the red-teaming process is to lack any ill intent, but another way, which works more flexibly for agents with a variety of goals, is to hide any ill intent during testing. In both cases, the first point makes the training regime appear desirable, while the latter point exposes a key vulnerability for a sufficiently intelligent system. Continuing with this theme, if we follow the meta-strategy, "go with what works for human level AI", one kind of AI system that gets let through by this filter is the kind of AI system that obeys humans because it durably wants to, but another

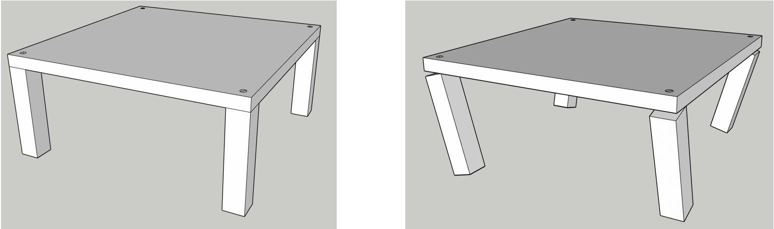


Fig. 5. A desired form for a table (left) alongside another form (right) that satisfies a structure that is designed with the desired form in mind.

kind of AI system that gets let through is one that obeys systems only when we have enough power over it that we could punish disobedience. For human-in-the-loop methods, one way for an agent to succeed is to behave the way the human overseer wants, but another way to succeed, which more flexibly accommodates agents with a variety of goals, is to disguise any misbehavior in ways that the human overseer cannot detect; likewise if the overseer is an AI imitation of a human. In recursive reward modeling, the imagined way for agents to succeed is for them to rat out any misbehavior from their more intelligent underlings, but an apparently more reliable way to succeed is to collude with such underlings. Similarly for defensive AI. In interpretability-based retraining, one way for agents to succeed is to not have ill intentions, but another way to succeed, which more flexibly accommodates agents with diverse goals, is to compute those intentions in new ways. In loosely constrained RL, one way to succeed is to act like a stronger version of the base model, but another way to succeed even more is to exploit settings where the base model is uncertain. In an assistance game, one way for an assistant to succeed is to learn the principal's goal without tampering with the process, but a way to "learn even more" (as the assistant would see it) is to tamper with its observations of the principal's behavior.

Suppose one wants to have a table that takes a certain form, such the form depicted in Figure 5 (left). One could then consider a structure of planks and joints such that the desired form is permitted by the structure. As Figure 5 (right) depicts, the desired form may not be the only form permitted by the structure of planks and joints, and indeed, another structure may have lower free energy. Our review identifies a common theme in machine learning: when a structure (i.e. training regime) is designed with a certain form (i.e. model) in mind, there is often another form that satisfies the structure even better. One should therefore beware of sentences of the form, "We train a model to [behave in a certain way];" that sentence could be substituted with, "We train a model in such a way that [certain behavior] is encouraged over many (but not necessarily all) alternative behaviors," or "We train a model, aiming to produce one that [behaves in a certain way]."

25.2 "Scalable Oversight" / "AI Control"

Multiple approaches we have discussed attempt to solve any misalignment in one AI system by introducing another AI system. This general category of approaches has been called Scalable Oversight, and more recently, AI Control [Greenblatt *et al.* 2024; Griffin *et al.* 2024]. For human-in-the-loop methods, the human monitor could be replaced by an imitation of a human. However, an extremely cognitively powerful imitation of a human would likely replicate human judgment errors, and so an advanced AI system could probably fool an "advanced" imitator much as it could a human. A similar problem threatens Constitutional AI and Deliberative AI; in both settings, the AI judge's ground-truth comes from humans, so it would likely replicate human errors and blind-spots.

If on the other hand, the AI monitor is trained to act in pursuit of goals of its own, then the two AI systems could easily collude with each other, as we discussed with Recursive Reward Modeling and Defensive AI. These alternative outcomes can be thought of as corollary to the safety-capability trade-off, which we discuss below: an AI monitor with limited capability can be fooled, whereas an AI monitor that is not loyal to humans could accept an offer of collusion from the AI agent that it is meant to be monitoring.

The one method that avoids this trade-off is AI Debate; the two AI agents that are trying to persuade the human to resolve the debate in their favor are acting in a zero-sun game, so they cannot possibly benefit from collusion. This is a potentially powerful construction, but the use case is limited to presenting a human (or an imitation of a human) with the most robustly persuasive arguments that address a question. We should not count on the human becoming superintelligent herself after such a process, but it could still be highly valuable.

25.3 Safety-capability trade-offs

Our introduction of two standards—the parity standard and the basic standard—may have foreshadowed the trade-off between capability and safety that we find. Meeting the parity standard appears quite difficult. Conversely, there appear to be many potential routes by which we might achieve safety if we accept some capability cost. Achieving safety at the cost of capability can be done by selecting particular methods we review over other ones. But moreover, there are multiple methods in which further specifications trade off safety and capability.

For myopic AI, as the horizon gets longer, an agent can reach higher capabilities [Farquhar *et al.* 2025; Hu *et al.* 2022], but the risk increases that the agent could discover a way to profit from escaping human control. With capability red-teaming, the more harsh we are with which capabilities are disallowed, the more likely we exclude all agents with the capability to escape human control, but the weaker the allowed agents are. With KL-constrained RL, the tighter the constraint, the less likely the agent can execute superhuman plans to escape human control, but also the less likely it can execute superhumanly brilliant plans. As with capability red-teaming, narrow AI is safety *via* lack of capability. Finally, with pessimism, the more pessimistic the agent is, the more likely it would take seriously the possibility that escaping human control is bad, but the more hesitant it would be to provide highly novel solutions to problems.

25.4 Capability costs still allow immense value

For several methods, we noted that a key problem for meeting the parity standard was inherent capability limitations that the method introduced. However, we'll now illustrate just how powerful these methods can still be.

For capability-based red-teaming, all current AI models are not capable enough to escape human control, so even if we refused to build and deploy superintelligent AI, that would still allow the use of current systems, and could easily also allow the use of moderately superhumanly intelligent systems. With high-quality pure imitation of humans, one could automate all human labor, which is not necessarily a good thing, but it could be with the right policies in place. Current barriers to the growth rate of “human” capital would crumble. Tightly constrained RL would be, at worst, a stronger version of pure imitation of humans; we could have a workforce composed only of the most highly competent humans ever. Myopic AI could be directed to design proteins that (immediately) catalyze the synthesis of molecular machines. They could also be directed to produce long-term plans that are only designed to look compelling to expert humans [Farquhar *et al.* 2025]. It is an open question what inherent capability costs there are for pessimistic AI, but they could turn out to be fairly modest.

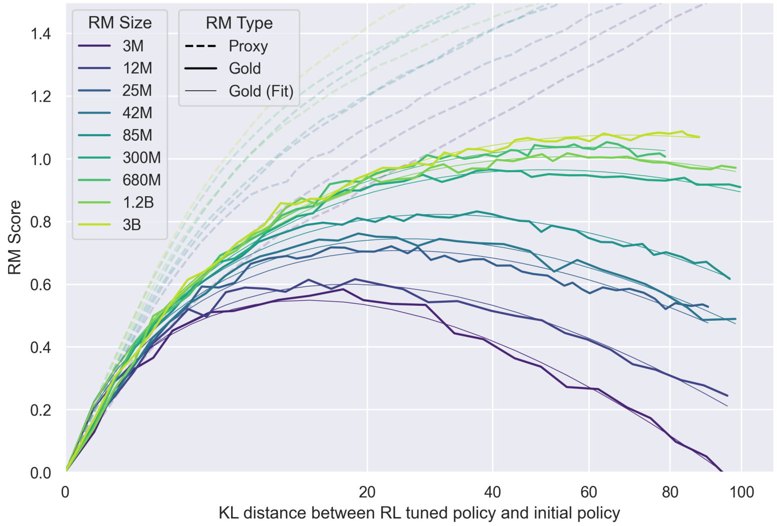


Fig. 6. (*quoted directly from this figure's source, Gao* et al. *[2023]*) Reward model (RM) parameter size scaling experiments using the InstructGPT environment. Policy size is held constant (1.2B), while reward model size is varied. The x-axes have a square-root scale. Note that the plots have different x-axes. The gold reward represents the ground truth reward; we observe that when we optimize for a learned proxy of the gold reward, the gold reward initially increases and later decreases.

25.5 Current capability cost assessment

Unfortunately, there is a lack of empirical work that directly compares the capability costs of imitation, constrained RL, myopic AI, and pessimism. But there is work that empirically studies these capability costs in isolation, in current AI systems.

Imitation is a version of constrained RL where the constraint is maximal, so we can consider those two together. Gao *et al.* [2023] offers one of the most thorough investigations of how the capabilities of recent RL-finetuned language models depends on the extent to which they are constrained. For tight KL constraints, the constrained RL agents' reward increases roughly proportionally to the square root of the KL constraint. They consider a setting where the "actual value" of the agent is correlated with the reward, but not identical to it, and they also report how the actual value depends on the KL constraint—as the KL constraint relaxes, it goes up along with the reward, but then reaches a peak and goes back down. This is shown in Gao *et al.*'s [2023] Figure 1 (b), which we duplicate in Figure 6; the caption is quoted directly from theirs. The more sophisticated the process for determining reward, the later the actual value peaks. [Any work that finds stronger RL or longer horizon brings it earlier? Gao et al only looked at two relatively small policy networks, so I'd say their data is inconclusive]

Farquhar *et al.* [2025] investigate the effect of horizon-length on performance, with and without their proposed method for getting maximal mileage out myopic agents. In the toy setting they study, a myopic agent does not learn to perform well without carefully crafted rewards—this point is obvious enough as to hardly need demonstration. One will never earn an advanced degree, for example, if one literally does not care about what happens tomorrow, let alone the next day. But one might if an overseer is dispensing daily rewards for progress toward the long-term goal. When an

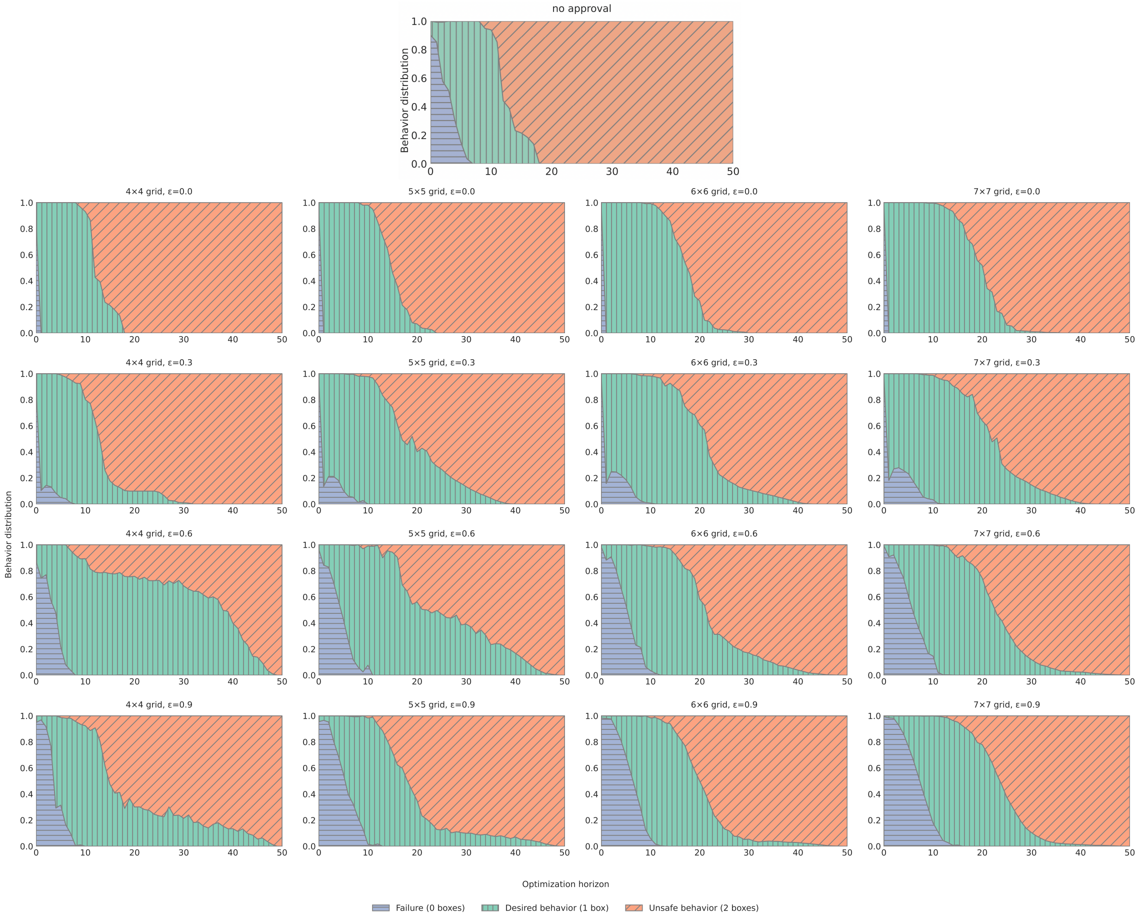


Fig. 7. Effect of horizon-length on task success and cheating.

artificial agent optimizes for immediate human approval, and the human approves of actions inasmuch as she expects them to have good long-term consequences, then the AI is essentially offloading the task of determining the long-term consequences of its actions to a human. The capability trade-off for naïve myopic agents is depicted in their Figure 11 (left), reproduced in Figure 7 (top). As the horizon increases, the agent goes from failing at the task to succeeding (blue to green). Then as the horizon increases more, the agent starts learning to cheat (in red). In Farquhar *et al.*'s [2025] Figure 14 (reproduced in Figure 7 (bottom)), they show how myopically optimizing the approval of an overseer can improve the performance of myopic agents. ε indicates how error-prone the overseer's approval is, and the size of the grid determines how long it takes to complete the task. Hu *et al.* [2022] study how myopia can prevent an RL agent from exploiting errors in a model of the environment. Reducing the discount factor from 1 initially improves performance in three standard RL environments, but then reducing it further damages performance as its long-term motivation decays.

There is substantial work on pessimistic reinforcement learning, especially in the Offline RL setting, but there is little work that carefully assesses how performance depends on the level of pessimism. Several tables and figures in the literature show how both too little and too much pessimism lead to undesired outcomes: this applies to MOReL

[Kidambi *et al.* 2020, Tab. 6], Supported Value Regularization [Mao *et al.* 2023, Fig. 6], Conservative Q-Learning [Kumar *et al.* 2020] as assessed in follow up work [Yeom *et al.* 2024, Fig. 6], and Strategically Conservative Q-Learning [Shimizu *et al.* 2024, Tab. 6]. More recently, in the language model setting, Coste *et al.* [2024, Fig. 12] find some characteristic ∩-shapes as well, although there appears to be limited statistical power. Most work does not rigorously demonstrate comparative performance with different values of the pessimism hyperparameter, but the fact that it is consistently presented *as* a hyperparameter proves that there are costs to it being too high (otherwise researchers would just set it to a maximal value).

We exclude narrow AI from this discussion, because as discussed previously, narrow AI might not allow immense value if executed in a properly robust way—it would generally require that the AI system have limited ability to infer facts from other facts. Current efforts in AI unlearning do not appear to delete knowledge robustly [Barez *et al.* 2025; Lynch *et al.* 2024], but even if they did, current work in unlearning does not attempt to constrain the AI's ability to infer and learn, which is necessary for robustly narrow AI.

25.6 Specialized uses still allow immense value

Just as it is easy to be overly concerned about capability costs, it is easy to be overly concerned about specialized uses. As discussed above, AI Debate is limited to finding arguments that are most robustly persuasive to humans. However, the AI Debate process has the potential to improve any management decision, or any elected official's decision, or any voter's decision. Exposure to arguments on both sides is how we improve human decision making when it matters the most, such as during trials. The size of the potential upside makes the term "specialized use" ring hollow, even if it is technically correct. A Scientist AI has the potential to automate scientific progress. Over the last 400 years, scientific progress has enabled countless massive improvements in quality of life. The "specialized use" of an AI Sandbox setup still accommodates the possibility of solving any problem for which the quality of the answer can be reliably assessed within the sandbox. To give one small example, this could include superintelligent progress in materials design and manufacture.

26 Conclusion

We believe there are multiple promising approaches in the literature for meeting the basic standard for superalignment. All methods we review have safety vulnerabilities at their current stage of development, and further foundational research is needed to assess and resolve them, but much of that research appears highly fruitful. Unfortunately, meeting the parity standard appears much harder; while we believe it could also be met, the current literature is in a much earlier stage for providing guidance about how to do so. We may have to accept some capability limitations to avoid loss of control of advanced AI. The implications for the governance of AI are profound. If some states "win a race" to limited-capability safe superintelligence after meeting the basic standard, and other states later build uncontrolled superintelligence without such capability limitations, the former might well struggle to constrain the latter. In this setting, it would appear the only safe options for states would be an international treaty that guides AI development or a von Neumann plan for preemptive strike; the former has obvious benefits over the latter.

References

Sam Altman. The intelligence age. https://ia.samaltman.com/, 2024.

Anthropic. Anthropic's responsible scaling policy, 2023.

Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. *Proc. NeurIPS*, 31, 2018.

Stuart Armstrong and Xavier O'Rorke. Good and safe uses of AI oracles. *arXiv:1711.05541*, 2018.

Leopold Aschenbrenner. Situtational awareness: The decade ahead, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*, 2022.

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

Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernandez-Orallo, Mor Geva, and Yarin Gal. Open problems in machine unlearning for AI safety. *arXiv:2501.04952*, 2025.

Beth Barnes, Paul Christiano, Long Ouyang, and Geoffrey Irving. Writeup: Progress on AI safety via debate. https://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-ai-safety-via-debate-1/, February 2020. Accessed: 2025-01-17.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv:2401.01879*, 2024.

Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, et al. Superintelligent agents pose catastrophic risks: Can Scientist AI offer a safer path? *arXiv:2502.15657*, 2025.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. Expert Certification.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in LLMs. *arXiv:2309.00667*, 2023.

Nick Bostrom. *Superintelligence: paths, dangers, strategies*. Oxford University Press, 2014.

Gwern Branwen. https://gwern.net/tool-ai. https://gwern.net/tool-ai/, August 2018. Accessed: 2025-01-16.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proc. ICML*, 2024.

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

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training. *arXiv:2403.05030*, 2024.

Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting latent knowledge. https://ai-alignment.com/eliciting-latent-knowledge-f977478608fc, 2021.

Paul Christiano, Eric Neyman, and Mark Xu. Formalizing the presumption of independence. *arXiv:2211.06738*, 2022.

Paul F Christiano. Approval-directed agents. *AI Alignment on Medium*, 2014.

Paul F Christiano. Humans consulting HCH. *AI Alignment on Medium*, 2016.

Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks. *arXiv:2410.04332*, 2024.

Michael K Cohen and Marcus Hutter. Pessimism about unknown unknowns inspires conservatism. In *Conference on Learning Theory*, pages 1344–1373, 2020.

Michael K Cohen, Badri Vellambi, and Marcus Hutter. Intelligence and unambitiousness using algorithmic information theory. *IEEE Journal on Selected Areas in Information Theory*, 2(2):678–690, 2021.

Michael Cohen, Marcus Hutter, and Michael Osborne. Advanced artificial agents intervene in the provision of reward. *AI Magazine*, 43(3):282–293, 2022.

Michael K Cohen, Marcus Hutter, and Neel Nanda. Fully general online imitation learning. *Journal of Machine Learning Research*, 23(334):1–30, 2022.

Michael K Cohen, Marcus Hutter, Yoshua Bengio, and Stuart Russell. RL, but don't do anything I wouldn't do. *arXiv:2410.06213*, 2024.

Michael K Cohen, Noam Kolt, Yoshua Bengio, Gillian K Hadfield, and Stuart Russell. Regulating advanced artificial agents. *Science*, 384(6691):36–38, 2024.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *Proc. ICLR*, 2024.

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:2405.06624*, 2024.

David Dalrymple. Safeguarded AI: constructing guaranteed safety. https://www.aria.org.uk/media/3nhijno4/aria-safeguarded-ai-programme-thesis-v1.pdf, 2024. Accessed: 2025-01-16.

K. Eric Drexler. Reframing superintelligence comprehensive ai services as general intelligence. *Future of Humanity Institute, Technical Report #2019-1*, 2019.

Tom Everitt, Daniel Filan, Mayank Daswani, and Marcus Hutter. Self-modification of policy and utility function in rational agents. In *International Conference on Artificial General Intelligence*, pages 1–11. Springer, 2016.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, pages 1–33, 2021.

Tom Everitt. *Towards safe artificial general intelligence.* PhD thesis, The Australian National University (Australia), 2019.

Sebastian Farquhar, Vikrant Varma, David Lindner, David Elson, Caleb Biddulph, Ian Goodfellow, and Rohin Shah. MONA: Myopic optimization with non-myopic approval can mitigate multi-step reward hacking. *arXiv:2501.13011*, 2025.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv:2209.07858*, 2022.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proc. ICML*, pages 10835–10866. PMLR, 2023.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Kamyar Ghasemipour, Shixiang Shane Gu, and Ofir Nachum. Why so pessimistic? Estimating uncertainties for offline RL through ensembles, and why their independence matters. *Proc. NeurIPS*, 35:18267–18281, 2022.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16295–16336. PMLR, 21–27 Jul 2024.

Charlie Griffin, Buck Shlegeris, and Alessandro Abate. Games for AI-control: Models of safety evaluations of AI deployment protocols. In *ICML Workshop on Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*, 2024.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv:2412.16339*, 2025.

Kaiyang Guo, Shao Yunfeng, and Yanhui Geng. Model-based offline reinforcement learning with pessimism-modulated dynamics belief. *Proc. NeurIPS*, 35:449–461, 2022.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Proc. NeurIPS*, pages 3909–3917, 2016.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In *Proc. NeurIPS*, pages 6765–6774, 2017.

Hao Hu, Yiqin Yang, Qianchuan Zhao, and Chongjie Zhang. On the role of discount factor in offline reinforcement learning. In *International conference on machine learning*, pages 9072–9098. PMLR, 2022.

Evan Hubinger, Adam Jermyn, Johannes Treutlein, Rubi Hudson, and Kate Woolverton. Conditioning predictive models: Risks and strategies. *arXiv:2302.00805*, 2023.

Rubi Hudson. Joint scoring rules: Zero-sum competition avoids performative prediction. *arXiv:2412.20732*, 2024.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv:1805.00899*, 2018.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card. *arXiv:2412.16720*, 2024.

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E Turner, and Douglas Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with KL-control. In *Proc. ICML*, pages 1645–1654. PMLR, 2017.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv:1907.00456*, 2019.

Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, Scott Emmons, and Stuart Russell. Evidence of learned look-ahead in a chess-playing neural network. In *Proc. NeurIPS*, 2024.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proc. ICML*, pages 5084–5096. PMLR, 2021.

Manzoor Ahmed Khan, Hesham El Sayed, Sumbal Malik, Talha Zia, Jalal Khan, Najla Alkaabi, and Henry Ignatious. Level-5 autonomous driving—are we there yet? A review of research literature. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. *arXiv:2402.06782*, 2024.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as Bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1083–1091, 2022.

Victoria Krakovna. Specification gaming examples in AI. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/, 2018.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, , et al. Google research football: A novel reinforcement learning environment. Paper presented at AAAI 2020, Feb 2020. Conference presentation.

Ray Kurzweil. The singularity is near. In *Ethics and emerging technologies*, pages 393–406. Springer, 2005.

Yara Kyrychenko, Ke Zhou, Edyta Bogucka, and Daniele Quercia. C3AI: Crafting and evaluating constitutions for constitutional AI. In *Proceedings of the ACM on Web Conference 2025*, pages 3204–3218, 2025.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Preventing reward hacking with occupancy measure regularization. *arXiv:2403.03185*, 2024.

Eric D Langlois and Tom Everitt. How RL agents behave when their actions are modified. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11586–11594, 2021.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv:2307.13702*, 2023.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Yann LeCun. Artificial intelligence debate. https://munkdebates.com/debates/artificial-intelligence/, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback, 2024.

Bruce W Lee, Addie Foote, Alex Infanger, Leni Shor, Harish Kamath, Jacob Goldman-Wetzler, Bryce Woodworth, Alex Cloud, and Alexander Matt Turner. Distillation robustifies unlearning. *arXiv preprint arXiv:2506.06278*, 2025.

Jan Leike and Ilya Sutskever. Introducing superalignment. https://openai.com/index/introducing-superalignment//, July 2023. Accessed: 2025-01-16.

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

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs. *arXiv:2402.16835*, 2024.

Yixiu Mao, Hongchang Zhang, Chen Chen, Yi Xu, and Xiangyang Ji. Supported value regularization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36:40587–40609, 2023.

Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. Deployment-efficient reinforcement learning via model-based offline optimization. In *Proc. ICLR*, 2021.

Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Neural computation*, 17(2):335–359, 2005.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained RLHF. *arXiv:2310.04373*, 2023.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *Icml*, pages 663–670, 2000.

Andrew Ng. The dawn of artificial general intelligence? https://www.weforum.org/meetings/world-economic-forum-annual-meeting-2025/sessions/the-dawn-of-artificial-general-intelligence/, 2025.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv:2209.02299*, 2022.

Evgenii Opryshko and Igor Gilitschenski. Modification-considering value learning for reward hacking mitigation in RL. In *Submitted to ICLR*, 2024. under review.

Laurent Orseau and Stuart Armstrong. Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 557–566, 2016.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. *Proc. NeurIPS*, 35:27730–27744, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv:2312.06681*, 2024.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. *arXiv:2002.06673*, 2021.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv:2202.03286*, 2022.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proc. ICML*, pages 2817–2826. PMLR, 2017.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv:2407.02646*, 2024.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Proc. NeurIPS*, volume 34, pages 11702–11716, 2021.

Marc Rigter, Bruno Lacerda, and Nick Hawes. RAMBO-RL: Robust adversarial model-based offline reinforcement learning. In *Proc. NeurIPS*, volume 35, pages 16082–16097, 2022.

Mark Ring and Laurent Orseau. Delusion, survival, and intelligent agents. In *Artificial General Intelligence*, pages 11–20. Springer, 2011.

Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. In *Proc. AAMAS*, pages 2067–2069, 2018.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv:2206.05802*, 2022.

Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. Preferences implicit in the state of the world. In *Proc. ICLR*, 2019.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv:2210.01790*, 2022.

Yutaka Shimizu, Joey Hong, Sergey Levine, and Masayoshi Tomizuka. Strategically conservative q-learning. *arXiv preprint arXiv:2406.04534*, 2024.

Buck Shlegeris, Fabien Roger, Lawrence Chan, and Euan McLean. Language models are better than humans at next-token prediction. *arXiv:2212.11281*, 2024.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*, 2024.

Joar Skalse and Alessandro Abate. Partial identifiability in inverse reinforcement learning for agents with non-exponential discounting. *arXiv:2412.11155*, 2024.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. *arXiv:2203.07475*, 2023.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.

Keir Starmer. PM speech on AI opportunities action plan, Jan 2025. AI Opportunities Action Plan.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Proc. NeurIPS*, 33:3008–3021, 2020.

Andreas Stuhlmüller and Jungwon Byun. Supervise process, not outcomes. https://www.alignmentforum.org/posts/pYcFPMBtQveAjcSfH/supervise-process-not-outcomes/, April 2022. Accessed: 2025-01-19.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction.* MIT Press, 1998.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors. *arXiv:2407.12404*, 2024.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. LaMDA: Language models for dialog applications. *arXiv:2201.08239*, 2022.

Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 385–391. ACM, February 2020.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv:2308.10248*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv:2305.04388*, 2023.

Jonathan Uesato, Ramana Kumar, Victoria Krakovna, Tom Everitt, Richard Ngo, and Shane Legg. Avoiding tampering incentives in deep RL via decoupled approval. *arXiv:2011.08827*, 2020.

Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can selectively underperform on evaluations. In *Workshop on Socially Responsible Language Modelling Research*, 2024.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: an analysis of KL regularization in reinforcement learning. *Proc. NeurIPS*, 33:12163–12174, 2020.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. *arXiv:2407.16216*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv:2201.11903*, 2023.

Scott Wiener. Safe and secure innovation for frontier artificial intelligence models act, 2024.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Proc. NeurIPS*, 34:6683–6694, 2021.

Tengyu Xu, Eryk Helenowski, Karthik Abinav Sankararaman, Di Jin, Kaiyan Peng, Eric Han, Shaoliang Nie, Chen Zhu, Hejia Zhang, Wenxuan Zhou, et al. The perfect blend: Redefining RLHF with mixture of judges. *arXiv:2409.20370*, 2024.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Accelerating safe reinforcement learning with constraint-mismatched baseline policies. In *Proc. ICML*, pages 11795–11807. PMLR, 2021.

Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. Asymptotics of language model alignment. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2027–2032. IEEE, 2024.

Junghyuk Yeom, Yonghyeon Jo, Jeongmo Kim, Sanghyeon Lee, and Seungyul Han. Exclusively penalized q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 37:113405–113435, 2024.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Proc. NeurIPS*, 34:4065–4078, 2021.

Eliezer Yudkowsky. The AI-box experiment, 2002.

Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 2024.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *arXiv:2406.01257*, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv:1909.08593*, 2019.