AI Alignment Strategies from a Risk Perspective: Independent Safety Mechanisms or Shared Failures?

Leonard Dung[1] **Florian Mai**[2,3]
[1]Ruhr-Universität Bochum
[2]Rheinische Friedrich-Wilhelms-Universität Bonn
[3]Lamarr Institute for Machine Learning and Artificial Intelligence
leonard.dung@ruhr-uni-bochum.de fmai@uni-bonn.de

Abstract

AI alignment research aims to develop techniques to ensure that AI systems do not cause harm. However, every alignment technique has *failure modes*, which are conditions in which there is a non-negligible chance that the technique fails to provide safety. As a strategy for risk mitigation, the AI safety community has increasingly adopted a *defense-in-depth* framework: Conceding that there is no single technique which guarantees safety, defense-in-depth consists in having multiple redundant protections against safety failure, such that safety can be maintained even if some protections fail. However, the success of defense-in-depth depends on how (un)correlated failure modes are across alignment techniques. For example, if all techniques had the exact same failure modes, the defense-in-depth approach would provide no additional protection at all. In this paper, we analyze 7 representative alignment techniques and 7 failure modes to understand the extent to which they overlap. We then discuss our results' implications for understanding the current level of risk and how to prioritize AI alignment research in the future.

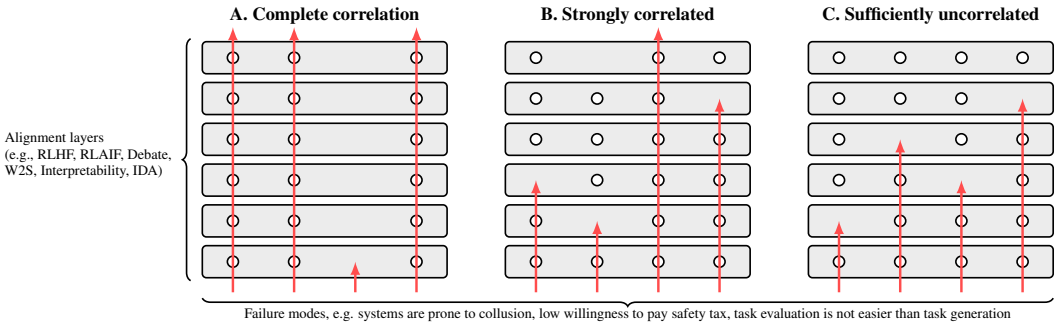

Figure 1: Correlation of failure modes governs defense–in–depth. **A**: Fully correlated presence/absence across layers (here, modes 1, 2, 4 succeed; mode 3 is universally absent and blocks at entry). **B**: Strong correlation leaves one shared mode (one straight–through attack), while others are blocked early. **C**: Sufficiently uncorrelated modes each block the attack at different layers.

1 Introduction

AI safety and alignment research aim to develop technical measures to ensure that AI systems do not cause harm, especially catastrophic harms through highly capable systems. We hold that any AI safety technique has *failure modes*, that is, conditions in which there is a non-negligible chance that it fails.

Preprint.

For example, certain interpretability-based techniques (e.g. representation engineering (Zou et al., 2023)) might fail if important safety-relevant behaviors are not linked to stable internal representations. Based on the defense-in-depth approach from the science of risk, our main claim is that a crucial question for AI safety research is to what extent the failure modes of different AI safety techniques are shared. Can different techniques be expected to fail in the same conditions? If yes, then AI risk is much higher than otherwise. In this case, special importance should be given to research on safety techniques which do not share the same failure modes. Our exploratory theoretical analysis tentatively concludes that many failure modes are shared between techniques, although there is also a notable degree of variation between the techniques we considered that give support the pursuit of certain alignment research directions.

In section 2, we provide an overview of a selected subset of alignment techniques. Section 3 describes the defense-in-depth approach to AI safety and highlights the resulting importance of correlations between failure modes. In section 4, we present a list of potential failure modes which we apply to the alignment techniques we reviewed, analyzing which failure modes may apply to which techniques. Section 5 outlines some implications of this failure mode analysis of AI safety as well as future research directions.

2 AI Alignment Techniques

2.1 Scope

The AI safety research field has experienced significant growth in recent years, spawning various AI alignment research directions and concrete technical contributions. In this section, we will review a collection of relevant techniques from the AI safety literature, focusing on techniques proposed to prevent catastrophic risks from highly capable general-purpose systems. *Forward alignment* methods (Ji et al., 2023) aim to achieve safety primarily through training and design interventions. *Backward alignment* includes monitoring, adversarial evaluation, and governance controls that mitigate harm even when systems are imperfectly aligned (e.g. Greenblatt et al. (2023)). We limit our scope to forward alignment families, set aside techniques for *learning under distribution shift* (Krueger et al., 2021) and safety evaluations (Phuong et al., 2024) and also abstract away from the integrated sociotechnical nature of safety (Bales, 2025; Friederich and Dung, 2025; Johnson and Verdicchio, 2025; Kasirzadeh, 2024) to keep the analysis focused. However, precisely since these approaches are importantly different from the ones we discuss here, they are promising elements in a comprehensive defense-in-depth approach. Within these confines, we aim to include paradigmatic examples of four categories of AI safety techniques that are currently highly influential, illustrating the broad applicability of failure mode analysis for AI safety.

2.2 Learning from Feedback

Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) uses reward models trained from human judgments over model outputs to steer the model toward desirable behavior. Its safety rationale is straightforward: If human raters consistently reward harmless behavior, the training process shifts the probability mass toward safer outputs. However, when applied to an AI model that only went through a pretraining stage, which often displays erratic behavior and low instruction-following ability, RLHF for helpfulness also improves the usability of the model. Since it can be flexibly added on top of other types of fine-tuning and incurs only a moderate cost, RLHF is now a default ingredient in many LLM pipelines at current capability levels. At the same time, RLHF relies on several strong assumptions such as that human raters recognize good model performance (Saunders et al., 2022; Sharma et al., 2023) and that reward models faithfully represent rater preferences (Lambert et al., 2024).

Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022) replaces or supplements human judgments with AI-produced feedback that is grounded in a human-written "constitution"; a set of rules that codifies desired behavioral principles. Compared to human feedback, AI feedback promises scalability to harder tasks and more data, and improved performance (Lee et al., 2023). We call the AI under alignment training the *generator* and the AI that gives feedback the *evaluator*. RLAIF was demonstrated to work even when the evaluator and the generator are fine-tuned from the same starting point Lee et al. (2023).

2.3 Scalable Oversight

AI Debate Irving et al. (2018) trains systems to argue opposing sides of a (binary) question (e.g. "which output is better?") in front of a human judge, who then selects the winner of the debate. By rewarding the winner accordingly, the training mechanism reinforces the behavior of the winning model. The underlying assumption is that it is easier to argue for the truth than for a falsehood. The safety mechanism is comparative: two adversaries expose one another's errors, reducing the judge's cognitive burden. Under idealized assumptions, debate protocols admit truth-favoring equilibria with formal guarantees even when debating against much stronger opponents (Brown-Cohen et al., 2023).

Burns et al. (2023) follow a standard approach for learning from feedback, in which, however, the (by assumption) "weak" human provides direct training supervision to an AI which is more capable than the human at the supervised tasks ("strong") (obtained through e.g. pretraining). Although the weak supervision is necessarily noisy, the *Weak-to-Strong Generalization* (W2S) hypothesis states that weak supervision of a strong AI yields better performance than the weak supervisor itself. Burns et al. (2023) find empirical evidence for W2S, albeit not across all tasks. The effect strengthens with an additional training loss that encourages the strong AI to be biased toward its own predictions and with a recursive bootstrapping approach. Various papers aim to explain why W2S occurs, e.g. Shin et al. (2024) identify that the strong AI learns from simple learning patterns present in some examples that the weak supervisor also detects. Alignment through W2S is attractive for both capability and safety: If a weak system's oversight reliably bootstraps a stronger one, we avoid relying on human-only supervision in superhuman regimes.

Iterated Distillation and Amplification (Christiano et al., 2018) provides a broader template in which a human supervisor decomposes a complex task into easier subtasks for weaker AIs to solve, which amplifies the overall competence of the human supervisor. By distilling the increased human competence back into the weak AIs, this method yields an aligned stronger AI, spawning a virtuous circle. Alignment is ensured by the fact that any capability and behavior that arises in the stronger AI originates from human supervision[1].

2.4 Interpretability-based Interventions

Interpretability-based methods aim to increase human understanding of AI systems, often based on the internal mechanisms producing behavior. *Representation engineering* (RE) (Zou et al., 2023) aims to extract representations of concepts such as honesty, power-aversion and morality by presenting the model with respective stimuli. Through extraction of neural activity (i.e., the latent representations) during a stimulus, RE identifies directions/features in activation space correlated with external features. By modifying the representation, e.g. through contrastive vector pairs that subtract or add undesirable/desirable concepts (Zou et al., 2023) or by training circuit-breaker models (Zou et al., 2024) that teach the model to break representations in undesirable scenarios, control can be exerted on the behavior of the model. Beyond RE, a complementary line uses sparse autoencoders to decompose activations into high-level concepts, enabling a more granular analysis and potential interventions at the feature level (Shu et al., 2025). While RE/circuit-breaking provides practical, post-hoc control with minimal retraining, SAE-based methods aim for greater causal specificity, albeit with significant scaling and evaluation challenges today.

2.5 Safety by Design

The techniques mentioned in the previous sections are highly compatible with the current deep learning paradigm that all of the state-of-the-art general-purpose AI systems follow: They present variations of how to perform supervised or reinforcement learning in order to update the weights of the neural network that represent the model, or intervene at inference-time by controlling the activations of the neurons. Due to the nature of neural networks as highly non-linear statistical models of often chaotic processes such as human behavior, it is difficult to mathematically prove their safety. Since safety-by-design approaches aim to construct systems with principled safety guarantees, most approaches that aspire safety-by-design either do not use neural networks or make additional assumptions that deviate significantly from the standard recipe to training SOTA AI models. We

[1]It has been argued that AI Debate is a form of IDA since the adversarial process is a form of amplification Irving et al. (2018). However, for the purpose of this paper, we will view the human's direct role in the decomposition process as central to IDA, which makes it qualitatively different from AI Debate.

restrict our scope to Scientist AI (Bengio et al., 2025a) since it provides the most holistic framework we are aware of. We leave the analysis of other promising directions for future work, e.g. POST agents Thornley (2025), which are designed to not resist shutdowns.

Scientist AI (Bengio et al., 2025a) proposes a non-agentic alternative to generalist agents that typically emerge from reinforcement learning. Rather than a system that plans and acts in the world to pursue goals, the goal is to create a system that merely explains observations and answers questions with explicit uncertainty, combining a theory-forming world model with a question-answering inference engine. The safety bet is architectural: If we can channel capability growth into explanatory competence (theories, predictions, calibrated uncertainty) rather than open-ended agency, we reduce exposure to instrumentally convergent behaviors (deception, power-seeking) that are prominent threat models in agentic designs (e.g. (Bostrom, 2012; Carlsmith, 2022; Dung, 2025). The proposal remains within the neural-network paradigm but departs from the standard "pretraining→SFT→RLHF" recipe. In its long-term plan, the world model and inference machine are trained from scratch to approximate the Bayesian posterior (and posterior-predictive), with the probabilistic inference machinery potentially implemented via GFlowNet objectives (Bengio et al., 2021) to sample and weight candidate theories. In principle, a Scientist AI may both (i) accelerate safety-relevant science and (ii) serve as a guardrail to any agentic system (regardless of how it was built).

3 Correlated Failure Modes and AI Safety

3.1 AI Safety and Defense-in-Depth

AI alignment research has traditionally often been conceived as looking for the solution to the alignment problem (Christian, 2020). Call this the "unitary model of AI safety" because it tries to achieve safety by finding a single extremely dependable safety technique. A different way of thinking seems more prominent to us now. In the influential Swiss Cheese Model of safety (Reason (2000); reviewed in Larouzee and Le Coze (2020); Shabani et al. (2024)), protections against accidents are represented by layers of Swiss cheese. However, each layer has holes - representing conditions in which the protection fails. Accidents happen normally only if the holes of the different layers line up, that is, if all protections fail simultaneously. The Swiss cheese model incorporates central ideas from the defense-in-depth framework. Defense-in-depth consists in having multiple protections against safety failure, such that safety can be maintained even if some protections fail. An overview paper states that "[d]efense-in-depth is a widely applied safety principle in practically all safety-critical technological areas" (Holmberg, 2017). For example, it is widely acknowledged that nuclear safety requires not only the high reliability of normal operation of power plants, but also various surveillance mechanisms and emergency response strategies if safety failures occur nevertheless.

In a recent paper, outlining their approach to technical AI safety, Google Deepmind (Shah et al., 2025, p. 63) puts forward an array of safety techniques, explicitly appealing to the defense-in-depth framework in the context of misuse risk. A post by OpenAI titled "How we think about safety and alignment" lists defense-in-depth more prominently as one core principle of their approach[2]. In a post in the alignment forum, the interpretability researcher Nanda (2025) argues that interpretability is best thought of as one layer in a comprehensive defense-in-depth strategy for safety.

The defense-in-depth model is superior to the unitary model of AI safety because there is no single safety technique that guarantees sufficiently high safety. When considering catastrophic risks, which by definition threaten extraordinary harm, an extremely high level of safety is required. It is doubtful that a single safety technique can afford this level of protection.

AI alignment research is sometimes described as lacking established theories, methods, and concepts (Kirchner et al., 2022; Chin, 2022). The empirical track record of AI safety supports the view that known methods are not completely trustworthy. In the International AI Safety Report 2025, (Bengio et al., 2025b, p. 23) conclude that "there has been progress in training general-purpose AI models to function more safely, but no current method can reliably prevent even overtly unsafe outputs." For example, despite massive efforts, LLMs are still susceptible to certain jailbreaks (e.g. (Millière, 2025)). Also, Hubinger et al. (2024) showed that standard alignment techniques, namely RLHF, finetuning on helpful, harmless, and honest outputs and adversarial training, can be jointly

[2] `https://openai.com/safety/how-we-think-about-safety-alignment/` (last accessed: 01.07.2025).

insufficient to eliminate a "backdoor" which makes the model produce undesirable behavior if and only if its prompt contains a certain specific trigger. Arvan (2025) even argues that perfect alignment, and interpretability, are provably impossible[3]. Moreover, all safety techniques depend on fallible bottlenecks, such as humans writing safety code and physical devices that are susceptible to malfunction; this means that safety techniques cannot be arbitrarily reliable (Cappelen et al., 2025).

By stacking multiple safety techniques which are redundant in the sense that each one is meant to be sufficient to ensure safety, the probability of failure can be dramatically reduced. Suppose safety depends on a single safety technique which has a probability of failure of 0.01. Then, the total probability of safety failure is 0.01. By contrast, imagine you have 10 safety techniques where each is ten times as unreliable, i.e. each has a probability of failure of 0.1. Assuming that these probabilities are independent and that a safety failure requires that they all fail simultaneously, the probability of safety failure is 0.0000000001. This is why, in realistic contexts, extremely high safety typically requires defense-in-depth.

3.2 The Importance of Correlated Failure Modes

Yet, suppose that the failure probabilities of our previous ten different techniques are perfectly correlated. If each technique is sure to fail if and only if each of the others fails, then the total probability of safety failure remains 0.1. In this scenario, defense-in-depth provides no additional protection at all.

If AI safety relies on a defense-in-depth strategy, then a crucial question is which AI safety techniques share failure modes. To our knowledge, this question has not received any dedicated treatment before. We will provide a theoretical analysis of the correlations between the failure modes of different AI safety techniques in the next section. We note that this question, essentially having to do with whether different techniques would fail in new circumstances, is by its very nature uncertain. However, we show that some insights can be obtained and thus provide a proof-of-concept for the usefulness of this research direction.

Knowledge about shared failure modes allows us to estimate the probability of an AI-induced catastrophe. If all our AI safety techniques are highly correlated, then this probability is much higher than otherwise. This suggests that, all other things being equal, development and deployment of AI systems that are candidate sources of catastrophic risk should proceed more cautiously if all our AI safety techniques are highly correlated.

Moreover, in a defense-in-depth strategy, safety techniques have disproportionate value if their failure modes are not highly correlated with other safety techniques. Thus, research efforts should often prioritize such techniques over ones that are less failure-prone but correlate more strongly with other safety techniques. It is also crucial to consider whether different techniques are compatible in the first place. While some methods (RLHF, RLAIF, weak-to-strong generalization, and interpretability-based interventions) are largely compatible with each other and straightforward to combine with today's predominant pretraining→SFT→RLHF pipeline, where most of the capabilities are learned during pretraining and only elicited and streamlined during the alignment process, others (IDA, Debate) make the assumption that learning the capabilities and alignment are part of the same process. Furthermore, Scientist AI proposes novel architectures and training paradigms. While it is conceivable that it can be integrated with the former group and the pretraining→SFT→RLHF pipeline, substantial empirical research is needed to understand whether this holds in practice.

Defense-in-depth is usually understood as stacking multiple layers of protection at once. However, it is also useful to maintain alignment methods in reserve that can be employed when conditions change. For example, it is not yet clear whether and when AIs will develop a resistance to shutdown through the current training paradigm (for evidence, see (Schlatter et al., 2025)). Should this become recognizable as a major problem in AI systems, it would be a significant advantage to have the option of switching to a training paradigm where this is less likely, e.g. POST agents (Thornley et al., 2024).

[3]Note that a mathematical guarantee of safety does not guarantee the absence of failure modes in our sense. The relevant formalization of safety may be too narrow and thus allow for some catastrophic outcomes or provably safe systems may lack performance-competitiveness and thus not be used.

4 A Failure Mode Analysis of AI Safety Techniques

4.1 What are Failure Modes?

A failure mode is a condition in which there is a non-negligible chance that a technique fails to provide safety. If other relevant factors are present and no other technique provides safety, then this condition makes an AI catastrophe possible. General failure modes may be shared between a wide and diverse range of safety techniques. In particular, their concrete technical implementation may be very heterogenous. By contrast, specific failure modes are specific to a narrow family of safety techniques which tend to be similar on a concrete technical level. For example, the question whether it is easier to persuade humans of truths compared to falsities in debates is central for the viability of the debate approach, making the absence of this condition one of its failure modes. Given the defense-in-depth approach, general failure modes of safety techniques are much more concerning than specific failure modes, since the former may be highly correlated among safety techniques, increasing the risk of joint failure. We will focus on seven potential general failure modes. They are based on a critical analysis of the safety techniques that we reviewed previously, informed by the literature.

1. Low willingness or capability to pay a safety tax. (S-TAX) A "safety tax" is the (not necessarily monetary) cost that may be required to ensure that a system is safe[4]. The central cost is often a reduction in the performance of systems that one develops or deploys. High-performing systems are valuable and, more important from a pure safety perspective, the higher the safety tax, the higher the incentive for actors to not pay the tax and instead proceed unsafely. An example: 2010-level AI systems are incapable of causing catastrophic harm. Nevertheless, "let's only build 2010-level AI" is not a viable safety proposal, because its safety tax is prohibitively high: There are strong benefits to developing more capable systems than 2010-level AI, making it likely that many actors will continue to do so.

2. Extreme or discontinuous AI capability development. (CAP-DEV) Some techniques may provide safety up until a certain level of AI capabilities, but fail to do once this level is crossed, or when capability advances are very rapid or involve sudden, unexpected "jumps".

3. Strong deceptive alignment tendencies emerge early during model development. (DEC-AL) It has been hypothesized, with recent tentative empirical evidence in support (Greenblatt et al., 2024), that sufficiently capable AI systems may pretend to be aligned to human goals, even if they are not (Cotra, 2021; Carlsmith, 2023; Dung, 2023). This way, they could increase the probability that humans, for example, deploy them or increase their capabilities and then, once humans cannot prevent it anymore, cause catastrophic harm. Such deceptive alignment may undermine certain safety techniques. Thus, a condition in which deceptive alignment tends to be strongly present and emerge relatively early on during model training is a failure mode for such techniques.

4. Systems are prone to collusion. (COLL) It has been hypothesized that multiple AI systems may collude with each other to undermine safety techniques that depend on using certain systems to monitor, align, or control others (e.g. (Hammond et al., 2025)).

5. The conditions for emergent misalignment are produced either accidentally or intentionally. (EM-MIS) Betley et al. (2025) show that language models trained with standard alignment techniques can, by fine-tuning on the narrow task of producing insecure code, be induced to exhibit undesirable behavior in a wide range of domains (broad misalignment). Thus, certain safety techniques fail to prevent such emergent misalignment, making it a failure mode. Although this phenomenon is still new and far from well understood, Wang et al. (2025) demonstrate that EM-MIS is caused by internal representations that correlate with "evil" personas learned during pretraining.

6. Task evaluation is not substantially easier than task generation. (EVAL-DIFF) Many safety techniques depend in some form on human- or AI-feedback. Such techniques may be more viable if providing valuable or accurate feedback on a task is substantially easier than performing the task.

[4] `https://forum.effectivealtruism.org/topics/alignment-tax` (Accessed: 07.08.2025).

Alignment method	S-TAX	CAP-DEV	DEC-AL	COLL	EM-MIS	EVAL-DIFF	AL-GEN
RLHF	✓	✗	✗	✓	✗	✗	✗
RLAIF	✓	✗	✗	✗	✗	✗	✗
Weak-to-Strong (W2S)	✓	✗	✗	✗	✗	✗	✗
AI Debate	?	✗	✓	✗	✗	✗	✗
Representation Engineering	✓	✓	✗	✓	✓	✓	✗
Scientist AI	✗	?	✓	?	✓	✓	✓
Iterated Amplification	✗	✓	✓	✓	✓	✓	✗

Table 1: Alignment methods (rows) vs. general failure modes (columns) as discussed in Secs. 4.1–4.2. Green (✓) = method does *not* suffer from the failure mode; Red (✗) = method *does*; Yellow (?) = unclear/mixed per the current analysis.

7. Systems generalize from alignment training in dangerous ways. (AL-GEN) The factors that cause state-of-the-art ML models to generalize in certain ways from training examples to situations outside the training distribution are not well-understood. This includes training for aligned behavior. If systems have dangerous tendencies for such out-of-distribution generalization, for example learning to behave aligned in training but changing behavior rapidly outside of situations encountered in training (Di Langosco et al., 2022), this may undermine certain safety techniques.

In the following, we analyze, for each of the alignment techniques and each of the failure modes we reviewed, what potential general failure modes an alignment technique has. Crucially, our selection of failure modes is not exhaustive and this analysis is highly exploratory and uncertain, all to be revised in the light of future empirical insights. However, it provides a proof-of-concept for the feasibility of failure-mode analysis for AI safety.

4.2 AI Safety Techniques and Their Failure Modes

This section analyses to what extent each alignment technique discussed in Section 2 is prone to each failure mode (see Section 4.1). For quick reference, we provide a summary in Table 1.

RLHF. RLHF is a technique standardly used on current models and also commercially valuable, since it includes helpfulness training and decreases behavior, such as producing insults, that AI companies desire to avoid independently of a concern for catastrophic risk reduction. Thus, low willingness to pay an alignment tax is not a failure mode. While the speed or discontinuity of AI capability advances is not specially relevant, it has been noted that RLHF does not scale to (sufficiently) superhuman AI capacities (e.g. (Casper et al., 2023)), since humans cannot provide accurate feedback on performance that is sufficiently beyond their own skills. For example, if such a system makes some statement, humans may want to provide positive feedback if and only if the statement is true but find themselves unable to assess the truth of the statement. So, very high AI capabilities are a failure mode. If strong deceptive alignment tendencies emerge early in training, this could undermine the RLHF training process. For example, systems could pretend to be aligned to avoid being changed by the training process but then show undesired behavior during deployment (Carlsmith, 2023; Cotra, 2021; Greenblatt et al., 2024). Since RLHF only involves training signals directly generated from human feedback, collusion between AI systems is not an option. Since, the way RLHF is standardly performed, no small group of humans has substantial influence on the feedback process, AI-human collusion is also not a failure mode. Empirical demonstrations of EM-MIS involve RLHF-trained models (Betley et al., 2025). Thus, while RLHF may not directly cause EM-MIS, it does not inherently prevent it from happening in RLHF-trained models. Thus, EM-MIS is a failure mode. If evaluation of outputs is not substantially easier than generation in many domains, then this makes it likely that RLHF breaks down in many domains relatively early, before reaching capacities far beyond human level. Thus, this is a failure mode. Finally, the efficacy of RLHF also depends on how systems naturally generalize from alignment training. If, for example, systems generalize from honesty training such that they learn to pretend to be honest in training or to only be honest within distribution, but are dishonest in deployment or outside the training distribution, this may undermine RLHF.

RLAIF. For analogous reasons to RLHF, S-TAX is not a failure mode of RLAIF. For the same reasons, DEC-AL, EVAL-DIFF, and AL-GEN are failure modes of RLAIF. The remaining failure modes require a more nuanced analysis. Since RLAIF is not limited to direct human feedback, it is not obviously vulnerable to the development of sufficiently capable systems, as RLHF is. However, to the best of our knowledge, RLAIF has not yet been shown to work when the evaluator is substantially weaker than the generator. Hence, a sufficiently discontinuous capability jump from one AI generation to the next might cause the AI feedback to fail, making RLAIF prone to CAP-DEV. Since RLAIF involves AI feedback, there is a risk that the generator and the evaluator collude, thus making COLL a failure mode for RLAIF. The original study on emergent misalignment (Betley et al., 2025) was performed primarily on OpenAI models. It is not publicly known whether these models are trained through RLAIF, so it is an open question whether RLAIF can actively prevent EM-MIS. However, given that RLHF and RLAIF provide similar feedback that is embedded into the model through the same gradient-based algorithms, it seems plausible that EM-MIS from narrow fine-tuning occurs for RLAIF-trained models in the same way.

Debate. As far as we know, there has been no demonstration of the integration of debate into a highly generally capable system, such as a state-of-the-art LLM. Moreover, envisioned debate paradigms are financially relatively expensive compared to established alignment methods like RLAIF because a) they involve a lengthy back-and-forth process involving multiple AI agents and rounds, and b) ultimately rely on human judge to pick the winner. At the same time, while debate requires additions to the current paradigm of training frontier AI systems, it is plausible that debate can be integrated into state-of-the-art systems in a cost-competitive manner, e.g. by using reward models (i.e., evaluator AIs) as proxies for humans (Irving et al., 2018; Kenton et al., 2024), or demonstrating that small amounts of debate training suffice. Hence, there is considerable uncertainty as to whether S-TAX is a failure mode for Debate; additional research is needed. Debate is explicitly designed to not break down when systems reach superhuman levels. Nevertheless, since the process relies on a human judge, it is plausible that there is some capability level at which humans would not be able to follow the debate and make justified judgments. For instance, sometimes people analogize the relation between the intellect of superintelligence and humans to the relation between humans and gorillas (Russell, 2019). It seems clear that gorillas could not reliably judge the outputs of a debate between humans in any debate format, making CAP-DEV a plausible failure mode of debate. As demonstrated by the formal results that honest strategies succeed (Brown-Cohen et al., 2023), if successful, Debate can uncover that a system is deceptively aligned, ruling out DEC-AL. A central concern is that both debaters might collude against humans, thus COLL is a clear failure mode. Since AIs trained through Debate would still fundamentally operate within the deep learning paradigm (likely including pretraining), there is no compelling reason to believe that it can prevent EM-MIS. Since Debate relies on humans' ability to discern which of two systems with beyond-human abilities is truthful, it assumes that, in the relevant context, task evaluation is easier than generation. Since Debate assumes that systems generalize from honesty training to robustly being honest, it assumes that systems generalize from alignment training in certain ways, making AL-GEN a failure mode.

Weak-to-strong generalization. Unlike Debate, W2S has been demonstrated in highly capable state-of-the-art systems. Moreover, since it leverages the well-established supervised fine-tuning pipeline, W2S may be the most efficient way of aligning high-capability systems, given current methods. Thus, S-TAX is not a pressing failure mode. Moreover, W2S is designed to achieve safety on arbitrary capable systems. However, W2S relies on bootstrapping; if AI capabilities advance discontinuously in large jumps, then a current-generation system may be insufficiently capable to provide good feedback for a next-generation system. So, CAP-DEV is a central failure mode. Possibly, a system which is deceptively aligned might undermine the training process, for the same reason that deceptive alignment is a concern with RLHF. At the same time, a sufficiently good feedback process may prevent DEC-AL. Nevertheless, DEC-AL constitutes a potential failure mode. Since the process involves multiple AIs, COLL is a possible failure mode. Since it operates within the pretrain→fine-tune context, EM-MIS constitutes a failure mode, similar to RLAIF. EVAL-DIFF and AL-GEN are failure modes, for the same reasons mentioned with respect to RLHF.

Representation engineering. Representation engineering (RE) is a low-cost fine-tuning procedure, and there is no evidence that it causes significant sacrifices to performance. Since RE detects the relevant circuits of the neural network that drive a behavior, it will likely work against systems with arbitrarily high performance. The speed or discontinuity of AI advances is not particularly

relevant to the effectiveness of RE. RE is a technique that is applied after the AI capabilities are developed. Hence, if deceptive alignment emerges early, e.g. during pretraining, systems may be able to systematically trick the RE technique (or prevent the human from executing it in the first place). While sufficiently advanced knowledge of mechanistic interpretability may enable recognizing such deception, this cannot be guaranteed. Collusion is not relevant to RE. Since EM-MIS strongly correlates with internal representations of persona features (Wang et al., 2025), RE could likely be used to guard systems against EM-MIS. RE does not rely on assumptions about the relative difficulty of task generation vs. evaluation. Crucially, especially when RE is combined with model training, a key concern is that RE may train systems to deceive more effectively, rather than abolishing deception (compare (Korbak et al., 2025)), such that dangerous generalization from alignment training is a key failure mode.

Scientist AI. S-TAX is a central failure mode; for the foreseeable future, it is likely that AI systems trained in the proposed way will be substantially less capable than transformer models trained through the common pretraining→SFT→RLHF pipeline. Thus, even if Scientist AI is safe, decision makers have strong incentives to nevertheless rely on alternative systems. If agency is necessary for catastrophic AI risk and AI capability is independent of agency, then Scientist AI may scale to arbitrary high AI capabilities. However, this is not obvious. It may be that agency emerges naturally, perhaps even almost inevitably, in sufficiently capable systems. Also, sufficiently capable systems may be capable of causing catastrophic harm, even if they (initially) lack agency. For example, mere superintelligent chatbots could manipulate humans, seducing them to obtain access to additional tools, or be misused by humans (compare (Bostrom, 2014, p. 148)). Thus, CAP-DEV is a failure mode. Agency is arguably a precondition of deceptive alignment which is why DEC-AL is not a failure mode. Collusion between systems may be a failure mode, if the non-agentic AI is used as a control measure for another agentic AI. However, if the non-agency proposal succeeds in making individual systems safe, it plausibly also succeeds at preventing collusion, but this remains an open question. There is no particular reason to think that emergent misalignment concerns translate to systems trained in a very different paradigm, without pretraining and reinforcement learning. The non-agency proposal also does not rely on assumptions about the relative difficulty of task evaluation and generation and it does not rely on alignment training: instead, it aims to eliminate the root of the possibility of misalignment, systems being agentic.

IDA. To the best of our knowledge, there is no state-of-the-art system that was trained through IDA. This can be explained by the fact that the cost of task decomposition through a human for every task is very high. Moreover, there is plausibly some limit to the capability level one can reach through IDA (albeit higher than human-level) since task decomposition itself may become too complex for human intelligence. This incurs a significant S-TAX compared to other training techniques. Yet, since AI's can still reach superhuman level, they may be prone to deception (DEC-AL) by or collusion (COLL) among AIs working on different sub-tasks. Due to the human's central role in task decomposition, which effectively governs what skills an AI will learn, CAP-DEV is not a failure mode. Since EM-MIS is a phenomenon that is rooted in pretraining, but IDA iteratively teaches (aligned) capabilities without pretraining, it is not a failure mode. Similarly, task evaluation (EVAL-DIFF) is not a relevant concept in IDA. However, even though IDA models are trained in a controlled fashion, they are still fundamentally based on neural networks that might generalize in dangerous, unexpected ways.

5 Discussion

Our analysis reveals that many failure modes may plausibly be shared between many different safety techniques. Generally, these results paint a concerning picture with respect to catastrophic AI risk. If the failure modes of different safety techniques are highly correlated, then catastrophic AI risk is much higher than it may seem.

Based on Table 4.1, we may sub-divide forward alignment techniques into three categories: On the one hand, techniques that are easy to implement (i.e. have a low safety tax) such as RLHF, RLAIF, and W2S share almost all failure modes. This can be explained by the fact that they all rely on the established pretraining→SFT→RLHF pipeline, which makes them prone to CAP-DEV, DEC-AL, EM-MIS, and AL-GEN, and that their alignment technique relies on the assumption that evaluation is easier than generation. On the other hand, techniques with a high safety tax (Scientist AI, IDA) are not prone to these same failure modes.

Debate and Representation Engineering belong to a third category: They have a moderate safety tax and are otherwise complementary to each other. For example, while Debate is the only affordable alignment technique that is not especially vulnerable to DEC-AL, it is prone to COLL, EM-MIS, EVAL-DIFF, and AL-GEN. In contrast, RE is prone to DEC-AL but not to the others.

From these observations, we can conclude that relying only on our current alignment techniques like RLHF, RLAIF, and W2S carries high risk because they have many correlated failure modes. In addition, we can motivate the following alignment research agendas:

First, on paper, Scientist AI and IDA are promising strategies for AI safety because they have the lowest number of failure modes. Of course, this comes at a high safety tax, which private frontier AI labs will be unlikely to pay due to the risk of falling behind. However, publicly financed mission-based initiatives (Mazzucato, 2018) are not subject to the same pressures, so - if equipped with substantial resources - they may be better positioned to pursue these strategies (compare (Goldstein and Salib, 2025) and (Petropoulos et al., 2025)). Nonetheless, the amount of safety tax matters for choosing which directions to pursue. Among Scientist AI and IDA, the former appears more tractable to us, as it does not depend on human annotations and has arguably no obvious performance ceiling. In contrast, the IDA framework may need to be redesigned to overcome the limitation driving its safety tax (Mai et al., 2025). Finally, our observations suggest that increasing the willingness and capability to pay safety taxes is crucial - political action is central here (Dung and Hellrigel-Holderbaum, 2025; Hendrycks et al., 2025; Katzke and Futerman, 2024). Since CAP-DEV is potentially a failure mode for most of the techniques we reviewed, political levers to shape the trajectory of AI development may be enormously valuable too.

Second, the combination of AI Debate and RE prevents almost all failure modes, revealing a potentially large opportunity for developing well-aligned AI if these techniques are compatible. This suggests that there should be significantly more research on how to integrate Debate with the classical pretraining→SFT→RLHF pipeline; whether Debate works in practice is an open question (Buhl et al., 2025), let alone whether it suffices to include small amounts of debate training in the pre- or post-training process to prevent early deceptive behavior from emerging. This is important because an early and strong tendency for deceptive alignment may also be a potential failure mode for all of the techniques we reviewed, except Debate and Scientist AI. Due to their pre-existing infrastructure, frontier AI labs are well-positioned to perform this research. If Debate indeed helps, we can apply RE on top as an additional layer of safety. In line with Nanda (2025)'s comment referenced above, our tentative analysis supports the view that interpretability-based methods are an important component of a defense-in-depth approach to AI safety, since representation engineering - our representative of the interpretability family - has a different distribution of failure modes than other techniques.

Third, Table 1 reveals that generalization from alignment training remains one of the most pressing research areas in AI safety, since all but one alignment method are prone to it.

6 Summary

In this paper, we reviewed some alignment techniques, highlighted the safety implications of correlations between failure modes of such techniques, provided an exploratory theoretical analysis of such correlations for the alignment techniques we reviewed, and discussed their implications. Our conclusions should be treated with care since our analysis is exemplary and not exhaustive; we consider only a sample of alignment techniques and possible failure modes, and although we aimed for representativeness, the complexity of the issue of correlated failure modes cannot be resolved with a single paper. Nonetheless, we believe that our work demonstrates the utility of failure mode analysis. We propose that future research - when analyzing the failure modes (or "limitations") of a wider range of safety proposals - should consider to what extent these failure modes are shared with other proposals. In addition, dedicated empirical and theoretical research should analyze what the failure modes of techniques are and whether they are shared. Subsequently, building on more mature and evidence-backed failure mode analyses than what we provided, future research effort should preferentially be directed to safety techniques whose failure modes promise to be highly uncorrelated from leading techniques. Moreover, these analyses can be used to inform assessments of catastrophic risks from AI and thus support decisions about which AI development or deployment to allow and under which conditions.

Acknowledgements

This work has been partially supported by the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

Author Contribution

Leonard Dung developed the initial idea for the paper. Since then, Leonard Dung and Florian Mai closely collaborated on all aspects of the writing process.

References

Arvan, M. (2025). 'interpretability'and 'alignment'are fool's errands: a proof that controlling misaligned large language models is the best anyone can hope for. *AI & SOCIETY*, 40(5):3769–3784.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bales, A. (2025). A polycrisis threat model for ai. *AI & SOCIETY*, pages 1–13.

Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. (2021). Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394.

Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., et al. (2025a). Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., et al. (2025b). International ai safety report. *arXiv preprint arXiv:2501.17805*.

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. (2025). Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford.

Brown-Cohen, J., Irving, G., and Piliouras, G. (2023). Scalable ai safety via doubly-efficient debate. *arXiv preprint arXiv:2311.14125*.

Buhl, M. D., Pfau, J., Hilton, B., and Irving, G. (2025). An alignment safety case sketch based on debate. *arXiv preprint arXiv:2505.03989*.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., et al. (2023). Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Cappelen, H., Dever, J., and Hawthorne, J. (2025). Ai safety: a climb to armageddon? *Philosophical Studies*, pages 1–18.

Carlsmith, J. (2022). Is power-seeking ai an existential risk? *arXiv preprint arXiv:2206.13353*.

Carlsmith, J. (2023). Scheming ais: Will ais fake alignment during training in order to get power? *arXiv preprint arXiv:2311.08379*.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Chin, Z. S. (2022). A newcomer's guide to the technical AI safety field. LessWrong.

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Christiano, P., Shlegeris, B., and Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cotra, A. (2021). Why ai alignment could be hard with modern deep learning. *Cold Takes*.

Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., and Krueger, D. (2022). Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR.

Dung, L. (2023). Current cases of ai misalignment and their implications for future risks. *Synthese*, 202(5):138.

Dung, L. (2025). The argument for near-term human disempowerment through ai. *AI & SOCIETY*, 40(3):1195–1208.

Dung, L. and Hellrigel-Holderbaum, M. (2025). Against racing to agi: Cooperation, deterrence, and catastrophic risks. *arXiv preprint arXiv:2507.21839*.

Friederich, S. and Dung, L. (2025). Against the manhattan project framing of ai alignment. *Mind & Language*.

Goldstein, S. and Salib, P. (2025). Collaboration at the brink: International law for the ai arms race. *Available at SSRN 5369439*.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., et al. (2024). Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Greenblatt, R., Shlegeris, B., Sachan, K., and Roger, F. (2023). Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., et al. (2025). Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.

Hendrycks, D., Schmidt, E., and Wang, A. (2025). Superintelligence strategy: Expert version. *arXiv preprint arXiv:2503.05628*.

Holmberg, J.-E. (2017). Defense-in-depth. *Handbook of safety principles*, pages 42–62.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., et al. (2024). Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Irving, G., Christiano, P., and Amodei, D. (2018). Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Johnson, D. G. and Verdicchio, M. (2025). The sociotechnical entanglement of ai and values. *AI & SOCIETY*, 40(1):67–76.

Kasirzadeh, A. (2024). Two types of ai existential risk: decisive and accumulative. *arXiv preprint arXiv:2401.07836*.

Katzke, C. and Futerman, G. (2024). The manhattan trap: Why a race to artificial superintelligence is self-defeating. *arXiv preprint arXiv:2501.14749*.

Kenton, Z., Siegel, N., Kramár, J., Brown-Cohen, J., Albanie, S., Bulian, J., Agarwal, R., Lindner, D., Tang, Y., Goodman, N., et al. (2024). On scalable oversight with weak llms judging strong llms. *Advances in Neural Information Processing Systems*, 37:75229–75276.

Kirchner, J. H., Smith, L., Thibodeau, J., McDonell, K., and Reynolds, L. (2022). Researching alignment research: Unsupervised analysis. *arXiv preprint arXiv:2206.02841*.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., et al. (2025). Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. (2021). Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., et al. (2024). Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Larouzee, J. and Le Coze, J.-C. (2020). Good and bad reasons: The swiss cheese model and its critics. *Safety science*, 126:104660.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., et al. (2023). Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Mai, F., Kaczér, D., Corrêa, N. K., and Flek, L. (2025). Superalignment with dynamic human values. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.

Mazzucato, M. (2018). Mission-oriented innovation policies: challenges and opportunities. *Industrial and corporate change*, 27(5):803–815.

Millière, R. (2025). Normative conflicts and shallow ai alignment: R. millière. *Philosophical Studies*, pages 1–44.

Nanda, N. (2025). Interpretability will not reliably find deceptive AI. AI Alignment Forum (forum post).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Petropoulos, A., Pataki, B., Juijn, D., Janků, D., and Reddel, M. (2025). Building cern for ai: An institutional blueprint. `https://www.cfg.eu/building-cern-for-ai`. Centre for Future Generations (CFG), Brussels.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., et al. (2024). Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*.

Reason, J. (2000). Human error: models and management. *Bmj*, 320(7237):768–770.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., and Leike, J. (2022). Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Schlatter, J., Weinstein-Raun, B., and Ladish, J. (2025). Shutdown resistance in large language models. *arXiv preprint arXiv:2509.14260*.

Shabani, T., Jerie, S., and Shabani, T. (2024). A comprehensive review of the swiss cheese model in risk management. *Safety in extreme environments*, 6(1):43–57.

Shah, R., Irpan, A., Turner, A. M., Wang, A., Conmy, A., Lindner, D., Brown-Cohen, J., Ho, L., Nanda, N., Popa, R. A., et al. (2025). An approach to technical agi safety and security. *arXiv preprint arXiv:2504.01849*.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Shin, C., Cooper, J., and Sala, F. (2024). Weak-to-strong generalization through the data-centric lens. *arXiv preprint arXiv:2412.03881*.

Shu, D., Wu, X., Zhao, H., Rai, D., Yao, Z., Liu, N., and Du, M. (2025). A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.

Thornley, E. (2025). Shutdownable agents through post-agency. *arXiv preprint arXiv:2505.20203*.

Thornley, E., Roman, A., Ziakas, C., Ho, L., and Thomson, L. (2024). Towards shutdownable agents via stochastic choice. *arXiv preprint arXiv:2407.00805*.

Wang, M., la Tour, T. D., Watkins, O., Makelov, A., Chi, R. A., Miserendino, S., Heidecke, J., Patwardhan, T., and Mossing, D. (2025). Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. (2023). Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Zou, A., Phan, L., Wang, J., Duenas, D., Lin, M., Andriushchenko, M., Kolter, J. Z., Fredrikson, M., and Hendrycks, D. (2024). Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373.