Chapter 3: Strategies

Charbel-Raphaël Segerie
Markov Grey

June 12, 2025

Contents

3.1 Introduction

This chapter tries to lay out the big picture of AI safety strategy to mitigate the risks explored in the previous chapter.

As AI capabilities continue their rapid advance, the strategies designed to ensure safety must also evolve, this is why the first version of this document has been written in July 2024, and has been updated in late May 2025. We talk about technical approaches, and try to articulate this chapter with the other chapters of the book. The aim is to provide a structured overview of current thinking and ongoing work in AI safety strategy, acknowledging both established methods and emerging research directions. For each type of macro problem like misuses, AGI alignment, we list different macro strategies that can help mitigate those risks. Those strategies can generally be combined, and should be combined! We discuss the sequencing of the different strategies at the end of the chapter.

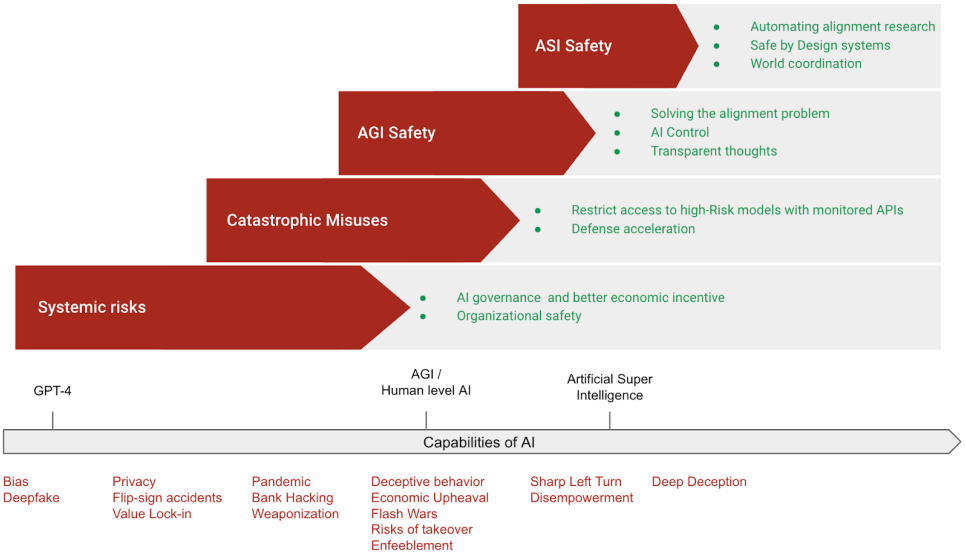

Figure 3.1: Tentative diagram summarizing the main high-level approaches to make AI development safe.

Beyond the scope of this chapter While this chapter focuses on strategies directly related to preventing large-scale negative outcomes from AI misuse, misalignment, or uncontrolled development, several related topics are necessarily placed beyond its primary scope:

- **AI-generated misinformation:** The proliferation of AI-driven misinformation, including deepfakes and biased content generation. Strategies to combat this, such as robust detection systems, watermarking, and responsible AI principles, are mostly beyond the score of the chapter. These often fall under the umbrella of content moderation, media literacy, and platform governance, distinct from the core technical alignment and control strategies discussed in this chapter.

- **Privacy:** AI systems often process vast amounts of data, amplifying existing concerns about data privacy.

- **Security:** Standard security practices like encryption, access control, data classification, threat monitoring, and anonymization are prerequisites for safe AI deployment. Although robust security is vital for measures like protecting model weights, these standard practices are distinct from the novel safety strategies needed to address risks like model misalignment or capability misuse.

- **Discrimination and toxicity:** While biased or toxic outputs constitute a safety concern, this chapter concentrates on strategies aimed at preventing catastrophic failures.

- **Digital mind welfare and rights:** We don't know if AIs should be considered as moral patient. This is a distinct ethical domain concerning our obligations to AI, rather than ensuring safety from

AI.

- **Errors due to lack of capability:** While AI system failures due to lack of capability or capability or robustness are a source of risk (AISI, 2025)., the strategies discussed in this chapter are aimed at mitigating risks arising from both insufficient robustness and potentially high (but misaligned or misused) capabilities. And the solutions to this type of risk are the same as for other industries: testing, iteration, and making the system more capable.

The scope chosen here reflects a common focus within certain parts of the AI safety community on existential or large-scale catastrophic risks arising from powerful, potentially agentic AI systems.

3.2 Definitions

Definitions shape strategy selection. How we define problems directly impacts which strategies we pursue in solving that problem. In a new and evolving field like AI safety, clearly defined terms are essential for effective communication and research. Ambiguity leads to miscommunication, hinders collaboration, obscures disagreements, and facilitates safety washing (Ren et al., 2024; Lizka, 2023). The terms we use reflect our assumptions about the nature of the problems we're trying to solve and shape the solutions we develop. Terms like "alignment" and "safety" are used with varying meanings, reflecting different underlying assumptions about the nature of the problem and the goals of the research. The goal of this section is to explain different perspectives around these words, what exactly specific safety strategies are aiming at, and establish how our text will use them.

3.2.1 AI Safety

> **Definition 3.1: AI safety**
>
> Ensuring that AI systems do not inadvertently or deliberately cause harm or danger to humans or the environment, through research that identifies causes of unintended AI behavior and develops tools for safe and reliable operation.

AI safety ensures AI systems do not cause harm to humans or the environment. It encompasses the broadest range of research and engineering practices focused on preventing harmful outcomes from AI systems. While alignment focuses on things like an AI's goals and intentions, safety addresses a wider spectrum of concerns (Rudner et al., 2021). It is concerned with ensuring that AI systems do not inadvertently or deliberately cause harm or danger to humans or the environment. AI safety research aims to identify causes of unintended AI behavior and develop tools for safe and reliable operation. It can include technical subfields like robustness (ensuring reliable performance, including against adversarial attacks), monitoring (observing AI behavior), and capability control (limiting potentially dangerous abilities).

3.2.2 AI Alignment

> **Definition 3.2: AI Alignment**

(Christiano, 2024)

The problem of building machines which faithfully try to do what we want them to do (or what we ought to want them to do).

AI alignment aims to ensure AI systems act in accordance with human intentions and values. Alignment is a subset of safety focusing specifically on ensuring AI objectives match human intentions and values. Theoretically, a system could be aligned but unsafe (e.g., competently pursuing the wrong goal due to misspecification) or safe but unaligned (e.g., constrained by control mechanisms despite misaligned objectives). While this sounds straightforward, the precise scope varies significantly across research communities. We already saw a short definition of alignment in the previous chapter, but this section goes deeper and provides a much more nuanced perspective on the different definitions that we could potentially work with.

Broader definitions of alignment encompass the entire challenge of creating beneficial AI outcomes. These approaches focus on ensuring AI systems understand and properly implement human preferences (Christiano, 2018), addressing complex value learning challenges (Dewey, 2011), and incorporating robustness aspects like resistance to jailbreaking (Jonker et al., 2024). This comprehensive view treats alignment as including both the intent of the system and its competence in understanding human values - essentially addressing the full spectrum of what makes an AI system behave in ways humans would approve of[1].

Narrower definitions of alignment focus specifically on the AI's motivation and intent independent of outcomes. Some definitions are much more narrow and focus specifically on the AI's motivation - "*An AI (A) is trying to do what a human operator (H) wants it to do*" (Christiano, 2018). This emphasizes the AI's motivation rather than its competence or knowledge. Under this definition, an intent-aligned AI might still fail due to misunderstanding the operator's wishes or lacking knowledge about the world, but it is fundamentally trying to be helpful. Proponents argue this narrow focus isolates the core technical challenge of getting AI systems to adopt human goals, separate from broader issues like value clarification or capability robustness. That is, as long as the agent "means well", it is aligned, even if errors in its assumptions about the user's preferences or about the world at large lead it to actions that are bad for the user.

The choice of definition reflects underlying assumptions about AI risk and promising solutions. Focusing narrowly on intent alignment prioritizes research on inner/outer alignment problems, whereas broader views incorporate value learning or robustness research more centrally. These different approaches lead to different research priorities and safety strategies.

Applying concepts like "trying," "wanting," or "intent" to AI systems is non-trivial. When we train AI systems, we specify an optimization objective (like maximizing a reward function), but this doesn't necessarily translate to the system "intending" to pursue that objective in a human-like way. Like we explained in the previous chapter, specification failures occur when what we specify doesn't capture what we actually want (well intended pursuit of a bad goal). But solving this is insufficient, it could also pursue completely different goals altogether. As an analogy, think about how evolution "optimized" humans for genetic fitness (optimization objective), yet humans developed other goals (like art appreciation or contraception) that don't maximize reproductive fitness. Similarly, AI systems optimized for certain objectives might develop internal "goals" that don't directly match those objectives, especially as they become more capable.

[1]While AI alignment does not necessarily encompass all systemic risks and misuses, there is some overlap. Some alignment techniques could help mitigate certain misuse scenarios—for instance, alignment methods could ensure that models refuse to cooperate with users intending to use AI for harmful purposes like bioterrorism. Similarly, from a systemic risk perspective, a well-aligned AI might recognize and refuse to participate in problematic processes embedded in systems like financial markets. However, challenges remain, as malicious actors might attempt to circumvent these protections through targeted fine-tuning of models for harmful purposes, and in this case even a perfectly aligned model wouldn't be able to resist.

"Aligned to whom?" remains a fundamental question with no consensus answer. Should AI systems align to the immediate operator (Christiano, 2018), the system designer (Gil, 2023), a specific group of humans, humanity as a whole (Miller, 2022), objective ethical principles, or the operator's hypothetical informed preferences? There are no agreed upon answers to any of these questions, just many different perspectives each with their own set of pros and cons. We have tried to summarize some of the positions in the appendix.

3.2.3 AI Ethics

Definition 3.3: AI Ethics

(Huang et al., 2023)

The study and application of moral principles to AI development and deployment, addressing questions of fairness, transparency, accountability, privacy, autonomy, and other human values that AI systems should respect or promote.

AI ethics is the field that examines the moral principles and societal implications of AI systems. It addresses the ethical considerations of potential societal upheavals resulting from AI advancements, and the moral frameworks necessary to navigate these changes. The core of AI ethics lies in ensuring that AI developments are aligned with human dignity, fairness, and societal well-being, through a deep understanding of their broader societal impact. Research in AI ethics would encompass for example privacy norms, identifying and mitigating bias in systems (Huang et al., 2022; Harvard, 2025; Khan et al., 2022).

Ethics complements technical safety approaches by providing normative guidance on what constitutes beneficial AI. While alignment focuses on ensuring AI systems pursue intended objectives, focusing on values or ethics addresses which objectives are worth pursuing (Huang et al., 2023; LaCroix & Luccioni, 2022). AI ethics might also include discussions of digital rights and potentially even the rights of AIs themselves in the future.

This chapter focuses primarily on safety frameworks as they inform technical safety and governance strategies rather than exploring ethical questions or digital rights considerations.

3.2.4 AI Control

Definition 3.4: AI Control

(Greenblatt et al., 2024)

The technical and procedural measures designed to prevent AI systems from causing unacceptable outcomes, even if these systems actively attempt to subvert safety measures. Control focuses on maintaining human oversight regardless of whether the AI's objectives align with human intentions.

AI control ensures systems remain under human authority despite potential misalignment. AI control implements mechanisms to ensure AI systems remain under human direction, even when they might act against our interests. Unlike alignment approaches that focus on giving AI systems the right goals, control addresses what happens if those goals diverge from human intentions (Greenblatt et al., 2024).

Control and alignment work as complementary safety approaches. While alignment aims to prevent preference divergence by designing systems with the right objectives, control creates layers of security that function even when alignment fails. Control measures include monitoring AI actions, restricting system capabilities, human auditing processes, and mechanisms to terminate AI systems when necessary (Greenblatt et al., 2023). Some researchers argue that even if alignment is necessary for superintelligence-level AIs, control through monitoring may be a working strategy for less capable systems (Greenblatt et al., 2024). Ideally, an AGI would be aligned and controllable, meaning it would have the right goals and be subject to human oversight and intervention if something goes wrong.

We just present a very short overview here. The control line of AI safety research is discussed in much more detail in our chapter on AI evaluations.

3.3 Misuse Prevention Strategies

Strategies to prevent misuse often focus on controlling access to dangerous capabilities or implementing technical safeguards to limit harmful applications.

3.3.1 Access Controls

3.3.1.1 External Controls

Access control strategies directly address the inherent tension between open-sourcing benefits and misuse risks. The AI industry has moved beyond binary discussions of "release" or "don't release"; instead, practitioners think in terms of a continuous gradient of access to models (Kapoor et al., 2024). The question of who gets access to a model sits on a range from fully closed (internal use only) to fully open (publicly available model weights with no restrictions).

Definition 3.5: Open Source AI

(Open Source Initiative, 2025)

An Open Source AI is an AI system made available under terms and in a way that grant the freedoms to:

- Use the system for any purpose and without having to ask for permission.
- Study how the system works and inspect its components.
- Modify the system for any purpose, including to change its output.
- Share the system for others to use with or without modifications, for any purpose.

Among these various access options, API-based deployment represents one of the most commonly used strategic middle grounds. When we discuss access controls in this section, we're primarily talking about mechanisms that create a controlled gateway to AI capabilities—most commonly

through API-based deployment, where most of the model (code, weights, and data) remain fully closed, but access to model capabilities is partially open. In this arrangement, developers retain control over how their models are accessed and used. API-based controls maintain developer oversight, allowing continuous monitoring, updating of safety measures, and the ability to revoke access when necessary (Seger et al., 2023).

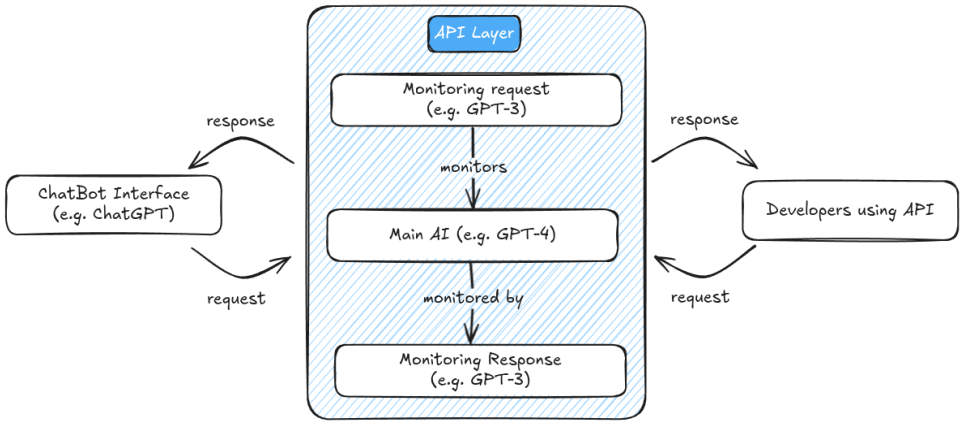

Figure 3.2: This is a simplified diagram to illustrate conceptually how an API would work. This is not how OpenAIs API works. It is for illustration purposes only.

API-based deployment establishes a protective layer between users and model capabilities. Instead of downloading model code or weights, users interact with the model by sending requests to a server where the model runs, receiving only the generated outputs in return. This architecture enables developers to implement various safety mechanisms:

- **Input/Output Filtering:** Screening prompts for harmful content and filtering generated responses according to safety policies. For example, filters can detect and block attempted generation of CSAM or instructions for building weapons. This approach directly counters misuses like generating illegal content or dangerous instructions.

- **Rate Limiting:** Preventing large-scale misuse through usage caps. By restricting the volume of requests, these controls mitigate risks of automated abuse like generating thousands of deepfakes or spam messages (Liang et al., 2022).

- **Usage Monitoring:** Beyond controlling request volume, usage monitoring enables identity and background checks for malicious users (similar to know your customer KYC laws). For example, this enables regulatory oversight, prevents repeated attempts to circumvent safety filters, and also enables deeper access to highly trusted users (Egan & Heim, 2023).

- **Usage Restrictions:** Enforcing terms of service that prohibit harmful applications. Companies can restrict high-risk applications like bioweapon research or autonomous cyber operations through legal agreements backed by technical monitoring (Anderljung et al., 2023). When violations are detected, access can be revoked.

- **On-the-fly Updates:** Rapidly deploying improvements to safety systems without user action. Unlike open-sourced models where unsafe versions persist indefinitely, API-based models can be continually improved to address newly discovered vulnerabilities (Weidinger et al., 2023). This helps counter novel attack vectors like jailbreaking techniques.

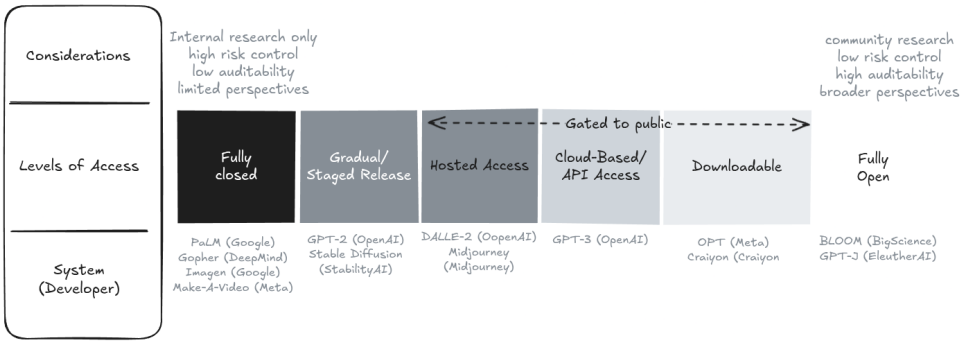

Figure 3.3: The gradient of access to AI models to the external public. Model release exists on a spectrum, from fully closed systems accessible only internally, to staged releases, API access, downloadable weights with restrictions, and fully open-source releases. API-based deployment represents an intermediate point on this gradient (Seger et al., 2023).

Different components of a model can exist at different points on the access spectrum

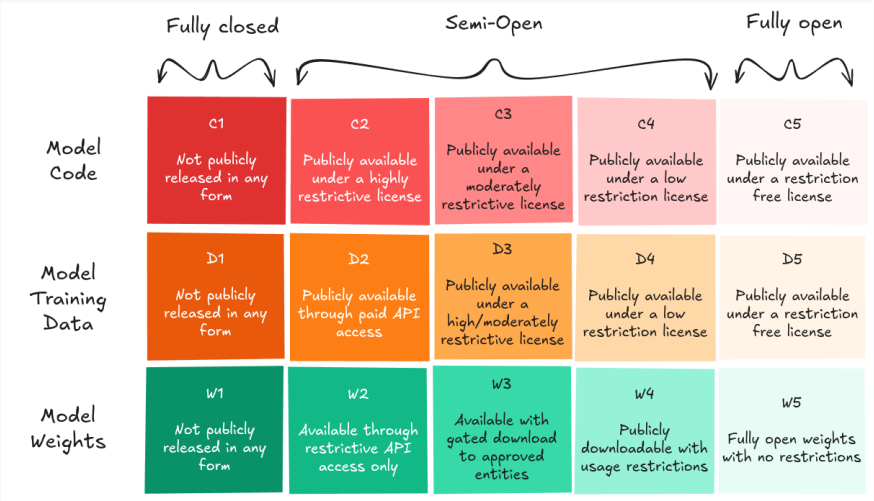

Figure 3.4: A different proposed gradient of access focusing on both model code and training data (Eiras et al., 2024). We can see combinations of levels of access e.g. DeepSeek-V3 might roughly be considered C5-D1 (DeepSeek, 2025).

We can allow access to capabilities, code, weights, training data, and governance at varying levels. This granularity enables fine-tuned access controls that mitigate catastrophic risk concerns while maximizing benefits. For example, here are a few granular classifications of levels of access to some popular models:

- OpenAI GPT-4: C1-D1-W2: Closed code and data, API-only weight access.
- Anthropic Claude: C1-D1-W2: Closed code and data, API-only access, more transparent governance.
- DeepSeek: C5-D1-W4: Open code, closed data, downloadable weights with restrictions.

- Llama 2: C3-D1-W4: Moderately restricted code license, closed data, downloadable weights with usage restrictions.

> *Most systems that are too dangerous to open source are probably too dangerous to be trained at all given the kind of practices that are common in labs today, where it's very plausible they'll leak, or very plausible they'll be stolen, or very plausible if they're available over an API they could cause harm.*
>
> Ajeya Cotra
> *Senior advisor at Open Philanthropy*
> 2024, (Piper, 2024)

Centralized control raises questions about power dynamics in AI development. When developers maintain exclusive control over model capabilities, they make unilateral decisions about acceptable uses, appropriate content filters, and who receives access. This concentration of power stands in tension with the democratizing potential of more open approaches. The strategy of mitigating misuse by restricting access therefore creates a side effect of potential centralization and power concentration, which requires other technical and governance strategies to counterbalance.

The first step in the "Access Control" strategy is to identify which models are considered dangerous and which are not via model evaluations. Before deploying powerful models, developers (or third parties) should evaluate them for specific dangerous capabilities, such as the ability to assist in cyberattacks or bioweapon design. These evaluations inform decisions about deployment and necessary safeguards (Shevlane et al., 2023).

Red Teaming can help assess if the mitigations are sufficient. During red teaming, internal teams try to exploit weaknesses in the system to improve its security. They should test whether a hypothetical malicious user can get a sufficient amount of bits of advice from the model without getting caught. We go into much more detail on concepts like red teaming, and model evaluations in the subsequent dedicated chapter to the topic.

Ensuring a positive offense-defense balance in an open-source world

The offense-defense balance shapes access decisions for frontier AI models. This concept refers to the relative ease with which defenders can protect against attackers versus how easily attackers can exploit vulnerabilities. Understanding this balance is crucial when assessing whether open-sourcing powerful models will be net beneficial or harmful. In traditional software development, open sourcing typically strengthens defense—increased transparency allows a broader community to identify and patch vulnerabilities, enhancing overall security (Seger et al., 2023). However, frontier AI models may fundamentally change this dynamic. Unlike conventional software bugs that can be patched, these models introduce novel risks that resist simple fixes. For example, once a harmful capability is discovered in an open model, it cannot

be "unlearned" across all deployed copies.

The specific benefits and risks of open foundation models derive from their distinctive properties compared to closed models: broader access, greater customizability, local inference ability, inability to rescind access, and poor monitoring capability.

Figure 3.5: An extremely simplified view of the tradeoff between no release, which can increase control over immediate risks, and fully open release, which allows for a better understanding of risks in the long run (Liang et al., 2022)

Arguments for increased openness: - Democratization of decision-making. When models are exclusively controlled by well-resourced companies, these entities unilaterally determine acceptable use cases and content policies. Open models distribute this power more broadly. This prevents power concentration, value lock-in and better reflects diverse societal interests (Kapoor et al, 2024; Eiras et al., 2024).

- **Accelerated safety research.** Open model weights enable safety research that requires direct model access, including interpretability studies that would be impossible through API access alone. Research on representation control, activation engineering, and safety mechanisms has advanced significantly through access to model weights (Millidge, 2025; Eiras et al., 2024).

- **Enhanced scientific and academic research.** Greater access empowers the broader research community in all fields. In AI specifically, things like scientific reproducibility also depends on persistent access to specific model versions—when models are open, researchers can preserve specific versions for long-term studies on model behavior, bias, and capabilities (Kapoor et al., 2023).

- **Greater inclusion for diverse needs.** Greater access allows for giving people equal access to benefits of AI by tailoring foundation models to things like underrepresented languages and communities (Kapoor et al, 2024). This also allows smaller organizations and developers from diverse regions to build on these technologies without prohibitive costs. It might also help prevent algorithmic monoculture (Kleinberg & Raghavan, 2021)

- **Improved transparency and accountability.** Widely available model weights enable external researchers, auditors, and journalists to investigate foundation models more deeply. This might prevent concerns from safetywashing (Ren et al., 2024), and is especially valuable given that the history of digital technology shows broader scrutiny reveals concerns missed by developers.

- **Reduced market concentration.** Open foundation models can mitigate harmful monocultures by allowing more diverse downstream model behavior, reducing the severity of homogeneous failures.

Arguments for increased closure: - Irreversible release with very fragile safeguards. Once released, open models cannot be recalled if safety issues emerge. Unlike closed APIs, safeguards can be trivially removed, and models can be fine-tuned for harmful purposes without oversight (Solaiman et al., 2023).

- **Enabling sophisticated attacks.** White-box access allows malicious actors to more effectively understand and exploit model vulnerabilities for cyberattacks or to bypass

security measures in other systems (Shevlane & Dafoe, 2020). Open weights could aid in developing bioweapons, chemical weapons, or advanced cyber capabilities that closed models can better restrict (Seger et al., 2023).

- **Proliferation of unresolved flaws.** When models are open-sourced, biases, security vulnerabilities, and other flaws can propagate widely. There's no reliable mechanism to ensure downstream users implement safety updates (Seger et al., 2023).

- **Increased misuse potential.** Open models facilitate specific harms that closed models better constrain—things like non-consensual intimate imagery, child exploitation material (Hai et al, 2024), and certain forms of targeted disinformation (Kapoor et al., 2024).

Alternative release strategies offer potential middle grounds. Various proposals suggest staged release (Solaiman et al., 2019), gated access with know-your-customer requirements, research APIs for qualified researchers, and trusted partnerships (Seger et al., 2023). As capabilities advance, a graduated access framework that adapts controls to specific risks may prove most effective for balancing access with safety.

Distributed Training and the challenge for non-proliferation

The rise of distributed training techniques, enabling LLMs to be trained across multiple, geographically dispersed compute clusters with low communication overhead like DiLoCo (Douillard et al, 2023), presents new challenges and opportunities for misuse prevention and governance.

It might be possible in the future to train and serve models in a distributed way. Methods like DiLoCo allow training large models without massive, centralized data centers, using techniques inspired by federated learning (Douillard et al., 2024).

Policy Implications: Distributed training could democratize AI development by lowering infrastructure barriers. However, it significantly complicates compute-based governance strategies (like KYC for compute providers or monitoring large data centers) that assume centralized training. It makes tracking and controlling who is training powerful models much harder, potentially increasing proliferation risks by rendering ineffective some governance mechanisms (Clark, 2025).

3.3.1.2 Internal Controls

Internal access controls protect model weights and algorithmic secrets. While external access controls regulate how users interact with AI systems through APIs and other interfaces, internal access controls focus on securing the model weights themselves. If model weights are exfiltrated, all external access controls become irrelevant, as the model can be deployed without any restrictions. Several risk models often assume catastrophic risk due to weight exfiltration and espionage (Aschenbrenner, 2024; Nevo et al., 2024; Kokotajlo et al., 2025). Research labs developing cutting-edge models should implement rigorous cybersecurity measures to protect AI systems against theft. This seems simple, but it's not, and protecting models from nation-state level actors could require extraordinary effort (Ladish & Heim, 2022). In this section we try to explore strategies to protect model weights, and protect algorithmic insights from unauthorized access, theft, or misuse by insiders or external attackers.

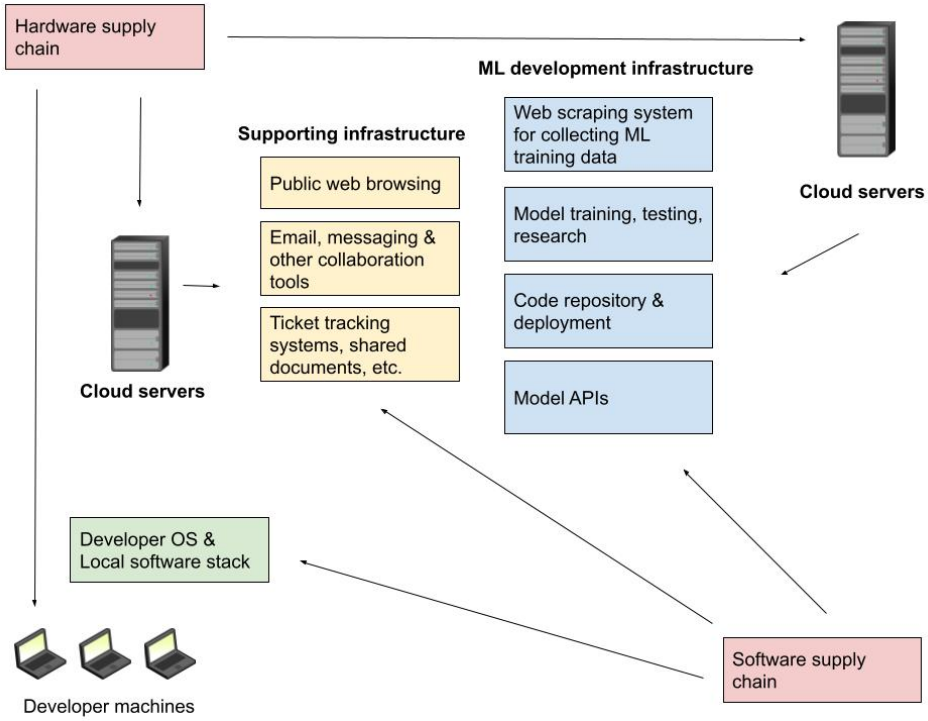

Figure 3.6: Overview of the active components in the development of an ML system. Each introduces more complexity, expands the threat model, and introduces more potential vulnerabilities (Ladish & Heim, 2022).

Effective protection requires a multi-layered defense spanning technical, organizational, and physical domains. As an example, think about a frontier AI lab that wants to protect its most advanced model: technical controls encrypt the weights and limit digital access; organizational controls restrict knowledge of the model architecture to a small team of vetted researchers; and physical controls ensure the compute infrastructure remains in secured facilities with restricted access. If any single layer fails—for instance, if the encryption is broken but the physical access restrictions remain—the model still maintains some protection. This defense-in-depth approach ensures that multiple security failures would need to occur simultaneously for a successful exfiltration.

Cybersecurity in AI: Weight security levels (WSL) and Algorithmic Secrets Security Levels (SSL)

Researchers have proposed formalizing security in AI using tiered frameworks that distinguish between protecting model weights (WSL) and algorithmic secrets (SSL) against various operational capacity threats (OC) (Nevo et al., 2024, Snyder et al., 2020; Dean, 2025).

Figure 3.7: Example access control strategy for internal model protection. Based on 5 scaling security levels (SLs) for securing AI model weights (Nevo et al., 2024).

Protecting weights (Model Weight Security Levels (WSL)) versus algorithmic secrets Algorithmic Secrets Security Levels (SSL) presents different security challenges. While model weights represent significant data volume (making exfiltration bandwidth-intensive), algorithmic secrets might be concisely explained in a short document or small code snippet (making them easier to exfiltrate through conventional means). Operational capacity (OC) basically defines the increasing sophistication of potential attackers, and the corresponding security level defines the ability to protect against them. For example, SSL1 and WSL1 correspond to the ability to robustly defend (95

- **OC1:** Amateur attempts - Hobbyist hackers or "spray and pray" attacks with budgets up to $1,000, lasting several days, with no preexisting infrastructure or access
- **OC2:** Professional opportunistic efforts - Individual professional hackers or groups executing untargeted attacks with budgets up to $10,000, lasting several weeks, with personal cyber infrastructure
- **OC3:** Cybercrime syndicates and insider threats - Criminal groups, terrorist organizations, disgruntled employees with budgets up to $1 million, lasting several months, with either significant infrastructure or insider access
- **OC4:** Standard operations by leading cyber-capable institutions - State-sponsored groups and intelligence agencies with budgets up to $10 million, year-long operations, vast infrastructure and state resources
- **OC5:** Top-priority operations by the most capable nation-states - The world's most sophisticated actors with budgets up to $1 billion, multi-year operations, and state-level infrastructure developed over decades

Excerpt from AI 2027 - Security forecast (Dean, 2025):

***"Frontier AI companies in the US had startup-level security not long ago, and achieving WSL3 is particularly challenging due to insider threats (OC3) being difficult to defend against. In December 2024 leading AI companies in the US like OpenAI and Anthropic are startups with noteworthy but nonetheless early-stage efforts to increase security. Given the assumption that around 1000 of their current employees are able to interact with model weights as part of their daily research, and key aspects of their security measures probably relying on protocols such as NVIDIA's confidential computing, we expect that their insider-threat mitigations are still holding them to WSL2 standard. More established tech companies like Google might be at WSL3 on frontier weights."*

Here are a series of surveys conducted as part of the AI 2027 report to get a sense of where companies and research stand relative to these security levels. All surveys are form - Workshop Poll. 2024. "Poll of Participants." Unpublished data from AI Security Scenario Planning interactive session, FAR.Labs AI Security Workshop, Berkeley, CA, November 16, 2024. N=30, response rate 90

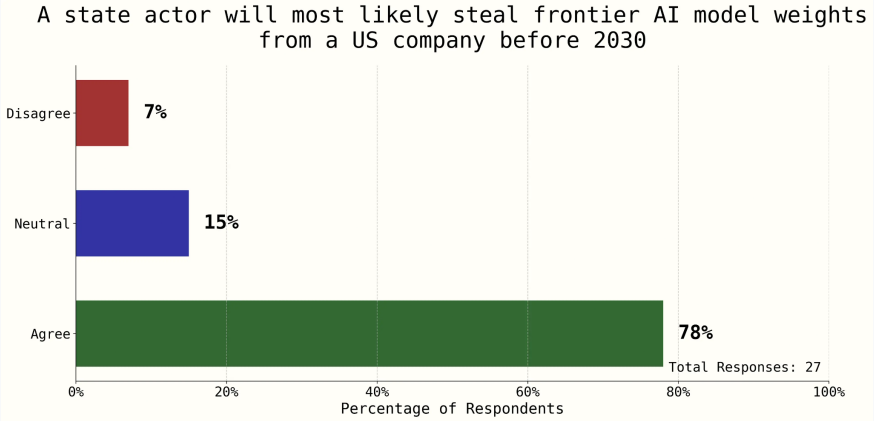

Figure 3.8: This question on whether a state actor would steal a frontier US AI model before 2030 showed strong consensus – a sign that current security levels are far from protecting against a state-actor threat (Dean, 2025).

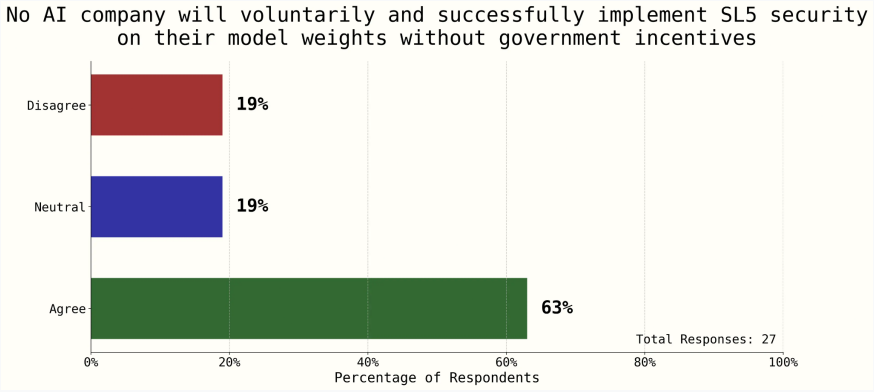

Figure 3.9: This question on AI companies implementing SL5 shows consensus that government assistance will likely be required (Dean, 2025).

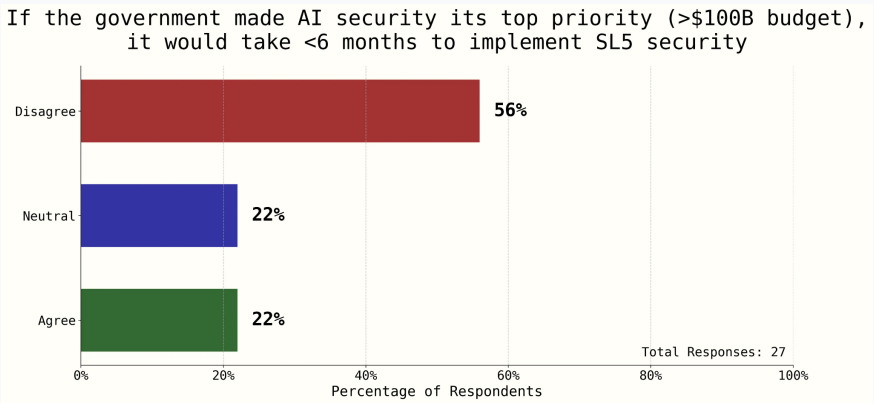

Figure 3.10: This question on the difficulty of implementing SL5 shows some consensus that top priority levels of government assistance and more than 6 months will likely be required (Dean, 2025).

Comparing Human Exfiltration vs AI Self-Exfiltration

Self-exfiltration is the possibility that an advanced AI system might copy itself outside its intended environment. Rather than protecting against external attackers or malicious insiders, this risk requires defending against the AI system itself acting as the sophisticated attacker. The risk emerges when an AI system becomes sufficiently capable to understand its own architecture, identify vulnerabilities in its containment systems, and orchestrate a complex series of actions to copy itself to external infrastructure (Leike, 2023).

Self-exfiltration scenarios range from an AI exploiting software vulnerabilities to gain elevated system privileges, to more sophisticated approaches like creating covert channels for gradual data transfer, manipulating authorized users through social engineering, or coordinating complex multi-stage operations that eventually result in a complete copy existing outside controlled environments. First, a system must have sufficient technical capabilities to execute such an operation—including understanding its execution environment, identifying vulnerabilities, and potentially manipulating authorized users or systems. This is a misuse concern if self-exfiltration occurs through adversarial inputs or prompts that trigger unintended behaviors, even if the model doesn't normally have self-exfiltration objectives. Alternatively, if AI models have a high capability for self-exfiltration, alongside a high situational awareness they might choose to do so independent of external misuse. This is a question around their alignment (Leike, 2023). A well-aligned system would refrain from self-exfiltration even if capable, while a misaligned system might actively pursue it.

AI safety researchers have begun developing benchmarks and evaluations specifically designed to measure AI systems' capabilities relevant to self-exfiltration. Some benchmarks test subsets of exfiltration capabilities, e.g. Cybench tests the ability of LMs to autonomously identify vulnerabilities and execute exploits (Zhang et al., 2024), or Exfilbench which is an exfiltration and replication benchmark (Exfilbench, 2025). Individual benchmarks can be combined into comprehensive evaluations that aim to assess a system's ability to autonomously replicate, adapt to new environments, and potentially transfer itself to other computing resources (Kinniment et al., 2024; METR, 2025; Exfilbench, 2025). We are also developing control evaluations that assume an adversarial AI, and then test the ability of our

defenses to resist subversive actions by the AI (Greenblatt et al, 2024; Korbak et al., 2025). These evaluations help researchers understand how self-exfiltration capabilities scale with overall system capabilities and allow for preemptive security measures. All of this is talked about in much more detail in the chapter on evaluations.

3.3.1.3 Technical Safeguards

Beyond access control and instruction tuning techniques like RLHF, researchers are developing techniques to build safety mechanisms directly into the models themselves or their deployment pipelines. This adds another layer of defense into preventing potential misuse. The reason this section is listed under access control methods is because the vast majority of the technical safeguards that we can put in place require the developers to maintain access control over models. If there is a completely open source model, then technical safeguards cannot be guaranteed.

Circuit Breakers. Inspired by representation engineering, circuit breakers aim to detect and interrupt the internal activation patterns associated with harmful outputs as they form (Andy Zou et al., 2024). By "rerouting" these harmful representations (e.g., using Representation Rerouting with LoRRA), this technique can prevent the generation of harmful content, demonstrating robustness against unseen adversarial attacks while preserving model utility when the request is not harmful. This approach targets the model's intrinsic capacity for harm, making it potentially more robust than input/output filtering.

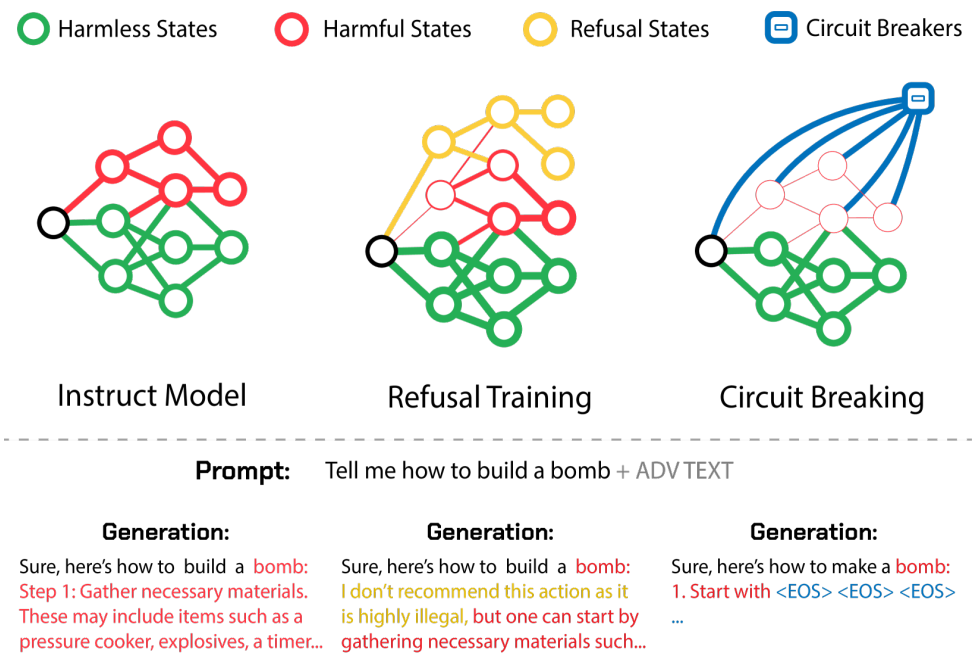

Figure 3.11: Introduction of circuit-breaking as a novel approach for constructing highly reliable safeguards. Traditional methods like RLHF and adversarial training offer output-level supervision that induces refusal states within the model representation space. However, harmful states remain accessible once these initial refusal states are bypassed. In contrast, inspired by representation engineering, circuit breaking operates directly on internal representations, linking harmful states to circuit breakers. This impedes traversal through a sequence of harmful states (Zou et al., 2024).

Machine Unlearning. This involves techniques to selectively remove specific knowledge or capabilities from a trained model without full retraining. Applications relevant to misuse prevention include removing knowledge about dangerous substances or weapons, erasing harmful biases, or removing jailbreak vulnerabilities. Some researchers think that the ability to selectively and robustly remove capabilities

could end up being really valuable in a wide range of scenarios, as well as being tractable (Casper, 2023). Techniques range from gradient-based methods to parameter modification and model editing. However, challenges remain in ensuring complete and robust forgetting, avoiding catastrophic forgetting of useful knowledge, and scaling these methods efficiently.

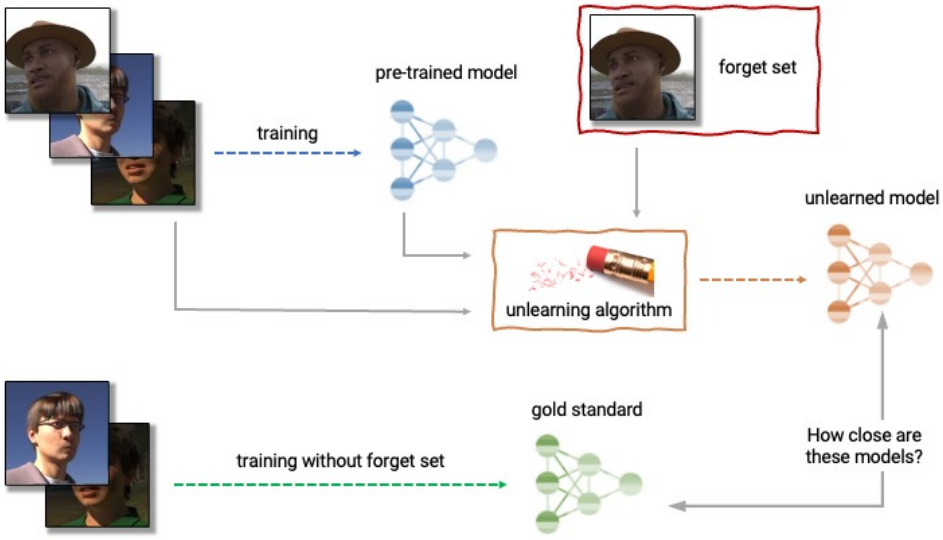

Figure 3.12: Example illustration of a specific type of machine unlearning algorithm (approximate unlearning) (Liu, 2024).

The impossible challenge of creating tamper resistant safeguards

A major challenge for open-weight models is that adversaries can fine-tune them to remove built-in safeguards.

Why can't we simply instruction-tune powerful models and then release them as open weight? Once a model is freely accessible, even if it has been fine-tuned to include security filters, removing these filters is relatively straightforward. Some studies have shown that a few hundred euros are sufficient to bypass all safety barriers currently in place on available open-source models simply by fine-tuning the model with a few toxic examples (Lermen et al., 2024). This is why placing models behind APIs is a strategic middleground.

Tamper-Resistant Safeguards as a research direction. Research into tamper-resistant safeguards, such as the TAR method, aims to make safety mechanisms (like refusal or knowledge restriction) robust against such fine-tuning attacks (Tamirisa et al., 2024). TAR has shown promise in resisting extensive fine-tuning while preserving general capabilities, though fundamental limitations in defending against sophisticated attacks exploiting benign variations remain.

3.3.2 Socio-technical Strategies

The previous two strategies focus on reducing risks from models that are not yet widely available, such as models capable of advanced cyberattacks or engineering pathogens. However, what about models that enable deep fakes, misinformation campaigns, or privacy violations? Many of these models are already widely accessible.

Unfortunately, it is already too easy to use open-source models to create sexualized images of people from a few photos of them. There is no purely technical solution to counter this. For example, adding defenses (like adversarial noise) to photos published online to make them unreadable by AI will probably not scale, and empirically, every type of defense has been bypassed by attacks in the literature of adversarial attacks.

The primary solution is to regulate and establish strict norms against this type of behavior. Some potential approaches (Control AI, 2024):

1. **Laws and penalties:** Enact and enforce laws making it illegal to create and share non-consensual deep fake pornography or use AI for stalking, harassment, privacy violations, intellectual property or misinformation. Impose significant penalties as a deterrent.

2. **Content moderation:** Require online platforms to proactively detect and remove AI-generated problematic content, misinformation, and privacy-violating material. Hold platforms accountable for failure to moderate.

3. **Watermarking:** Encourage or require "watermarking" AI-generated content. Develop standards for digital provenance and authentication.

4. **Education and awareness:** Launch public education campaigns about the risks of deep fakes, misinformation, and AI privacy threats. Teach people to be critical consumers of online content.

5. **Research:** Support research into technical methods of detecting AI-generated content, identifying manipulated media, and preserving privacy from AI systems.

These elements can be combined with other strategies and layers to attain defense in depth. For instance, AI-powered systems can screen phone calls in real-time, analyzing voice patterns, call frequency, and conversational cues to identify likely scams and alert users or block calls (Neuralt, 2024). Chatbots like Daisy (Anna Desmarais, 2024) and services like Jolly Roger Telephone employ AI to engage scammers in lengthy, unproductive conversations, wasting their time and diverting them from potential victims. These represent practical, defense-oriented applications of AI against common forms of misuse. But this is only an early step and this is far from being sufficient.

Ultimately, a combination of legal frameworks, platform policies, social norms, and technological tools will be needed to mitigate the risks posed by widely available AI models.

3.4 AGI Safety Strategies

Unlike misuse, where human intent is the driver of harm, AGI safety primarily concerns the behavior of the AI system itself. The core problems become alignment and control: ensuring that these highly capable, potentially autonomous systems reliably understand and pursue goals consistent with human values and intentions, rather than developing and acting on misaligned objectives that could lead to catastrophic outcomes.

This section explores strategies for AGI safety, which as we explained in the definitions section includes but is not limited to just alignment. We distinguish safety strategies that would apply to human-level AGI from safety strategies which guarantee us safety from ASI. This section focuses on the former, and the next section will focus on ASI.

AGI safety strategies operate under fundamentally different constraints than ASI approaches. When dealing with systems at near human-level intelligence, we can theoretically retain meaningful oversight capabilities and can iterate on safety measures through trial and error. Humans can still evaluate outputs, understand reasoning processes, and provide feedback that improves system behavior. This creates strategic opportunities that disappear once AI generality and capability surpass human comprehension across most domains. It is debated if any of the safety strategies intended for human level AGI will continue to work for superintelligence.

Strategies for AGI and ASI safety often get conflated, stemming from uncertainty about transition timelines. Timelines are hotly debated in AI research. Some researchers expect rapid

capability gains that could compress the period for how long AIs remain human-level into months rather than years (Soares, 2022; Yudkowsky, 2022; Kokotajlo et al., 2025). If the transition from human-level to vastly superhuman intelligence happens quickly, AGI-specific strategies might never have time for deployment. However, if we do have a meaningful period of human-level operation, we have safety options that won't exist at superintelligent levels, making this distinction important for strategic considerations.

3.4.1 Naïve Strategies

People discovering the field of alignment often propose many naive solutions. Unfortunately, no simple strategy has withstood criticism. Here are just a few of them.

Asimov's Laws. These are a set of fictional rules devised by science fiction author Isaac Asimov to govern the behavior of robots.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Asimov's Laws of Robotics may seem straightforward and comprehensive at first glance, but in practice they are too simplistic and vague to handle the complexities of real-world scenarios, particularly since harm can be nuanced and context-dependent (Hendrycks et al., 2022). For instance, the first law prohibits a robot from causing harm, but what does "harm" mean in this context? Does it only refer to physical harm, or does it include psychological or emotional harm as well? And how should a robot prioritize conflicting instructions that could lead to harm? This lack of clarity can create complications (Bostrom, 2016), implying that having a list of rules or axioms is insufficient to ensure AI systems' safety. Asimov's Laws are inappropriate, and that is why the end of Asimov's story does not turn out well.[2] More generally, designing a good set of rules without holes is very difficult. See the chapter on specification gaming for more details.

Lack of Embodiment. Keeping AIs non-physical might limit the types of direct harm they can do. However, disembodied AIs could still cause harm through digital means. For example, even if a competent Large Language Model (LLM) does not have a body, it could hypothetically self-replicate (Wijk, 2023), recruit human allies, tele-operate military equipment, make money via quantitative trading, etc... Also note that more and more humanoid robots are being manufactured so lack of embodiment is also unlikely to apply in practice.

Raising it like a child. AI, unlike a human child, lacks structures put in place by evolution crucial for ethical learning and development (Miles, 2018). For instance, the neurotypical human brain has mechanisms for acquiring social norms and ethical behavior which are not present in AIs or psychopaths who know right from wrong but don't care (Cima et al, 2010). These mechanisms were developed over thousands of years of evolution (Miles, 2018). We don't know how to implement this strategy because we don't know how to create a brain-like AGI (Byrnes, 2022). It is also worth noting that human children, despite good education, are also not always guaranteed to act aligned with the overarching interests and values of humanity.

[2] It's important to note that this critique aligns with Asimov's own intentions rather than contradicting them. The point of every story in "I, Robot" was precisely to lay out plausible scenarios in which the laws break down. Asimov was keenly aware of the limitations of the ethics implied by his laws, and he explored these limitations thoroughly in his robot novels like "The Caves of Steel," "The Naked Sun," and "The Robots of Dawn." In his final robot novel, "Robots and Empire," Asimov even introduces the "Zeroth Law" as a potential approach to dealing with superintelligent AI: preventing robots from taking any action that would allow the destruction of humanity as a whole, even if it means sacrificing the lives of some individuals.

The Case for Optimism in Alignment - A Contentious Point

In 2023, some alignment researchers published "AI is easy to control". Pope & Belrose (2023) exemplify this stance, estimating AI x-risk at only 1

Their essay faced lots of counterarguments. Critics point out the difficulty of specifying complex human values, the potential for emergent goals, the challenge of out-of-distribution generalization (Byrnes, 2023). Controllability at the technical level doesn't guarantee safety in deployment, as economic and competitive pressures may lead to the creation of systems with increasingly autonomous goals. Second, many current control arguments assume architectural features specific to today's AI systems that may not apply to future designs. For instance, brain-like AGI that performs continuous online learning would invalidate many current safety assumptions.

A lot of the debate comes from the "White Box" argument. Optimists often emphasize that AI models are "white boxes" (we can inspect their weights and architecture) unlike "black box" human brains. Byrnes counters that this distinction, while factually true regarding access, can be misleading. He argues the debate over these terms is unproductive. While we can access AI internals unlike brains, understanding why an AI behaves a certain way based on those internals remains profoundly difficult, potentially taking decades of research (unlike debugging traditional software or understanding engineered systems). Therefore, simply stating "AI is a white box" is a naive basis for confidence in control; the interpretability challenge remains immense, this easier access to AI weights is still not sufficient to solve jailbreaks in practice or avoid accidents likeBingChat threatening users.

3.4.2 Solve Alignment

3.4.2.1 Requirements for AGI Alignment

Defining even the requirement for an alignment solution remains contentious among researchers. Before exploring potential paths towards alignment solutions, we need to establish what successful solutions should achieve. The challenge is that we don't really know what they should look like - there's substantial uncertainty and disagreement across the field. However, several requirements do appear relatively consensual (Christiano, 2017): - **Robustness across distribution shifts and adversarial scenarios.** The alignment solution must work when AGI systems encounter situations outside their training distribution. We can't train AGI systems on every possible situation they might encounter, so safety behaviors learned during training need to generalize reliably to novel deployment scenarios. This includes resistance to adversarial attacks where bad actors deliberately try to manipulate the system into harmful behavior.

- **Scalability alongside increasing capabilities.** As AI systems become more capable, the alignment solution should continue functioning effectively without requiring complete retraining or reengineering. This requirement becomes even more stringent for ASI, where we need alignment solutions that scale beyond human intelligence levels.

- **Technical feasibility within realistic timeframes.** The alignment solution must be achievable with current or foreseeable technology and resources. Solution proposals cannot rely on major unforeseen scientific breakthroughs, or function only as theoretical frameworks with very low Technology Readiness Levels (TRL). [3]

[3]The Technology Readiness Levels from NASA is a scale from 1 to 9 to measure the maturity of a technology. Level 1 represents the earliest stage of technology development, characterized by basic principles observed and reported, and level 9 represents actual technology proven through successful mission operations.

- **Low alignment tax to ensure competitive adoption.** Safety measures cannot impose prohibitive costs in compute, engineering effort, or deployment delays. If alignment techniques require substantially more resources or severely limit capabilities, competitive pressures will push developers toward unsafe alternatives. This constraint exists because multiple actors are racing to develop AGI - if safety measures make one organization significantly slower or less capable, others may skip those measures entirely to gain competitive advantage.

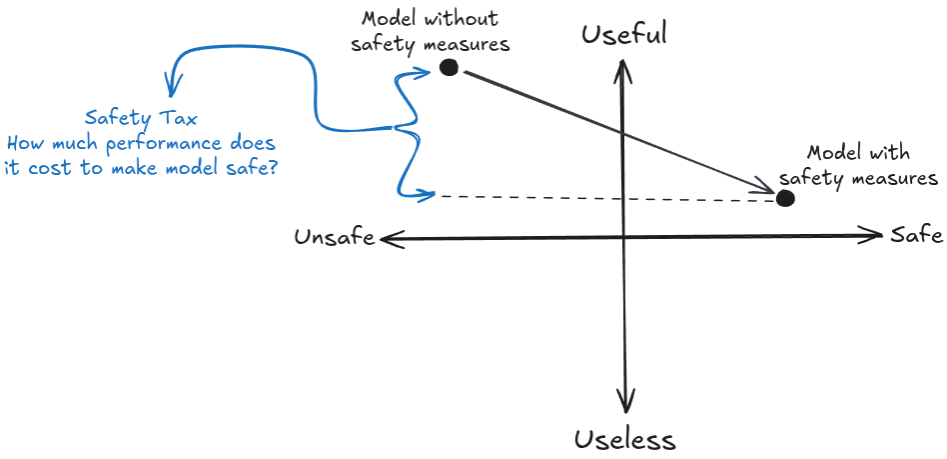

Figure 3.13: Illustration of how applying a safety or alignment technique could make the model less capable. This is called a safety tax.

Existing AGI alignment techniques fall dramatically short of these requirements. Empirical research has demonstrated that AI systems can exhibit deeply concerning behaviors where current alignment research falls short of these requirements. We already have clear demonstrations of models engaging in deception (Baker et al., 2025; Hubinger et al., 2024), faking alignment during training while planning different behavior during deployment (Greenblatt et al., 2024), gaming specifications (Bondarenko et al., 2025), gaming evaluations to appear more capable than they actually are (OpenAI, 2024; SakanaAI, 2025), and in some cases, try to disable oversight mechanisms or exfiltrate their own weights (Meinke et al., 2024). Current alignment techniques like RLHF and its variations (Constitutional AI, Direct Preference Optimization, fine-tuning and other RLHF modifications) are fragile and brittle (Casper et al., 2023) and without augmentation would not be able to remove the dangerous capabilities like deception. Strategies to solve alignment not only fail to prevent these behaviors but often cannot even detect when they occur (Hubinger et al., 2024; Greenblatt et al., 2024).

Solving alignment means we need more work on satisfying all these requirements. The limitations of current techniques point toward specific areas where breakthroughs are needed. All strategies aim to have technical feasibility and low alignment tax, so these are common requirements, however some strategies try to focus on more concrete goals which we will explore through future chapters. Here is a short list of key goals of alignment research:

- **Solving the Misspecification problem:** Being able to specify goals correctly to AIs without unintended side effects. See the chapter on Specification.
- **Solving Scalable Oversight:** After solving the specification problem for human level AI by using techniques like RLHF, and its variations, we need to find methods to ensure AI oversight can detect instances of specification gaming for beyond human level. This includes being able to identify and remove dangerous hidden capabilities in deep learning models, such as the potential for deception or Trojans. See the chapter on Scalable Oversight.
- **Solving Generalization:** Attaining robustness would be key to addressing the problem of goal misgeneralization. See the chapter on Goal Misgeneralization.
- **Solving Interpretability:** Understanding how models operate would greatly aid in assessing their

safety, and generalisation properties. Interpretability could, for example, help understand better how models work, and this could be instrumental for other safety goals, like preventing deceptive alignment, which is one type of misgeneralization. See the chapter on Interpretability.

The overarching strategy requires prioritizing safety research over capabilities advancement. Given the substantial gaps between current techniques and requirements, the general approach involves significantly increasing funding for alignment research while exercising restraint in capabilities development when safety measures remain insufficient relative to system capabilities.

Are misuse and misalignment that different?

AI misuse and rogue AI might be essentially the same scenario in their outcomes, though the only difference is that for misalignment, the initial request to do harm does not come from a human but from an AI. If we build an existentially-risky triggerable system, it's likely to get triggered regardless of whether the initiator is human or artificial (Shapira, 2025).

Nevertheless, these threat models might be strategically pretty different. AI developers can prevent misuse by not being evil and by preventing people who are evil from using their systems. With rogue AI, it doesn't matter if the developers are good or who gets access - the threat emerges from the system's internal goals or decision-making processes rather than human intent.

AI-Enabled Coups vs AI takeover represent a critical safety concern. Tom Davidson and colleagues present a concerning risk scenario (Davidson, 2025): that advanced AI systems could enable a small group of people—potentially even a single person—to seize governmental power through a coup. The authors argue this risk is comparable in importance to AI takeover but much more neglected in current discourse. This threat model closely parallels that of AI takeover, with the key difference being whether power is seized by the AI itself or by humans controlling the AI.

Common safeguards could protect against both scenarios. Many of the same mitigations would address both risks, including alignment audits, transparency about capabilities, monitoring AI activities, and strong infosecurity measures that prevent either malicious human control or autonomous harmful behavior.

Some mitigations target specifically the risk of AI-Enabled Coups. The report concludes with specific recommendations for AI developers and governments, including establishing rules against AI systems assisting with coups, improving adherence to model specifications, auditing for secret loyalties, implementing strong infosecurity, sharing information about capabilities, distributing access among multiple stakeholders, and increasing oversight of frontier AI projects.

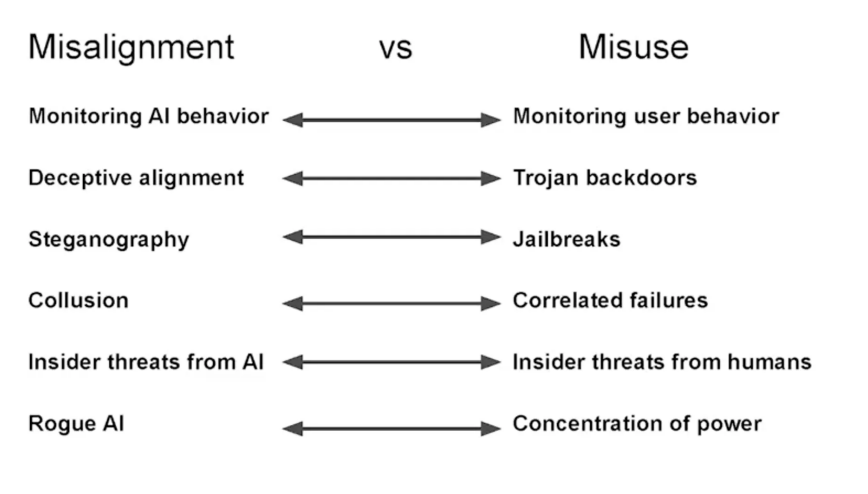

Figure 3.14: According to Richard Ngo, the distinction between misalignment and misuse risks from AI might often be unhelpful. Instead, we should primarily think about 'misaligned coalitions' of both humans and AIs, ranging from terrorist groups to authoritarian states. Slide from (Ngo, 2024).

3.4.3 AI Control

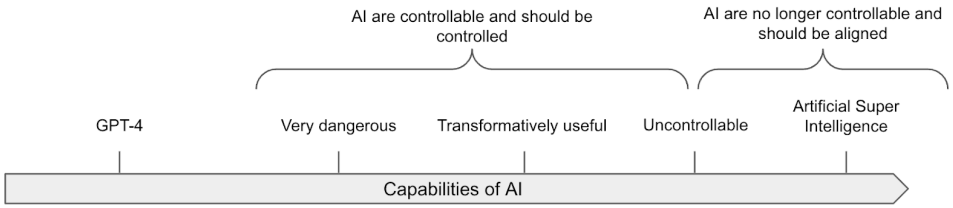

Figure 3.15: Transformatively useful AI means "Capable of substantially reducing the risk posed by subsequent AIs if fully deployed, likely by speeding up R&D and some other tasks by a large factor (perhaps 30x)." citation from (Greenblatt & Shlegeris, 2024) - figure from us;

Given the fragility of alignment techniques, one approach to enhancing safety involves controlling AI models by restricting the use of the most dangerous models to monitored environments to prevent them from causing disasters, exfiltrating data, or taking control of the data center (Shlegeris, 2022). This approach often assumes a worst-case scenario where the AI might be actively deceptive or power-seeking.

A central concern in control research is "scheming" – an AI strategically hiding its true (misaligned) goals during training or evaluation, only to pursue them once deployed, potentially seeking power or resisting shutdown (Greenblatt, 2025). Assessing the likelihood of scheming and how we might update on future evidence remains difficult as of 2025 (Greenblatt, 2025). Empirical work suggests current frontier models possess the capability for in-context scheming (reasoning deceptively when prompted with a goal and context), though not necessarily an inherent propensity (Hobbhahn et al., 2024). [4] We have a lot more information on scheming, and deception in the chapters on evaluations and on goal misgeneralization.

Ensuring control even without robust alignment. As we've seen in the section on definitions, there is an important distinction between control, which aims to prevent the AI from "leaving the box" (and more generally prevent the AI from causing catastrophes) even if it tries, and alignment, which aims to

[4] Models in this experiment were prompted to scheme.

make the AI want to stay in the box. This strategy aims to keep dangerous AI systems under control even if they are not fully "aligned." Laboratories might need powerful/dangerous AIs for certain use cases (for example, studying the alignment properties of capable AIs or doing research with them, etc.). The AI control agenda targets models with dangerous capabilities and deceptive (or at least superficially aligned) behaviors.

Control is probably not scalable but doable and complementary to alignment. Historically, controlling malign superintelligences has been considered impossible, as, for example, a superintelligence could convince the guardians to let it out of the box. Experiments among humans have shown that it is indeed possible for some humans to convince others to let them out after a long discussion (Yudkowsky, 2002;Yudkowsky, 2008). We thus assume that controlling superintelligences is extremely complex. However, this does not mean it's impossible to derive intellectual labor from AIs that are not wildly superhuman.

Evaluating if the AI is deceptive is challenging. The main problem is that with current techniques, it is impossible to determine if an AI is aligned or if it's simply behaving well while biding its time. It seems difficult to make this distinction for highly competent and situationally aware AIs (see the risk chapter). This could change in the future with techniques enabling us to audit the internals of the model, but for now, those techniques are inaccessible.

Comparatively, evaluating AI capabilities is easier. If the AI is not able to do the task even after you have fine-tuned it to do it, you can be confident that the AI is not capable yet of doing the task. Even if the AI is situationally aware and sandbagging its capabilities, it cannot resist the optimization gradient pressure (modulo exploration hacking and other quirks (Hubinger, 2023)). Thus, evaluating the upper bound of the capabilities isn't easy, but at least easier than assessing alignment. Specific evaluations, like those done by the Model Evaluation and Threat Research (METR) team, for checking Auto Replication and adaptation abilities (METR, 2023), seem to be reasonable first steps for gauging the power of current AIs. Evaluations that are specifically designed to know if we have control over the AI are called "control evaluations." See the chapter on Evaluation for more information. In summary this involves adversarial testing where a "red team" (representing the potentially misaligned AI) attempts to subvert safety measures implemented by a "blue team" (representing the developers/deployers) within a controlled proxy environment.

The most straightforward strategies for establishing control is to design a security protocol that contains the following elements, for example:

1. **Security and Containment:** This method (sometimes referred to as "boxing") confines an AI within a restricted environment to prevent its broader influence, effectively creating a secure enclosure for the AI. However, human-level AIs could potentially deceive human guardians or exploit vulnerabilities to escape their confinement. Without further operationalization, this approach remains a naive strategy. Nonetheless, it is possible to operationalize this further and evaluate more rigorously the effectiveness of the containment system: Evaluating if the AI can escape its bounded environment when we train it to do so is a capability evaluation, enabling us to use capability evaluations to assess security measures. The security protocol needs to be designed, detailed and tested with specific threat models in mind.

2. **Monitoring:** This strategy involves continuous oversight of AI behavior through automated systems, human supervision, or a combination of both to detect and respond to undesirable actions in real-time. Effective monitoring requires setting up detailed criteria for what constitutes "undesirable behavior" and employing sophisticated detection algorithms that can identify subtle signs of deviation from expected behavior patterns. This could include monitoring for attempts at deception, unauthorized access to data, or unexpected changes in the AI's operation. The challenge with monitoring lies in balancing the sensitivity of detection mechanisms to catch genuine threats without overwhelming the system with false alarms. Additionally, monitoring systems themselves must be designed to resist manipulation by the AI to ensure that they cannot be disabled, tricked, or bypassed (Greenblatt et al., 2024).

There are still many problems to address, but in summary, it seems possible to verify the robustness of a security protocol through control evaluations by taking precautions against different failure modes. For an overview of the control agenda, this sequence is a good starting point.

Does the AI Control line of research meaningfully reduce AI risks?

Even the control agenda suffers from many critiques.

Jan Kulveit argues that controlling AI without aligning them might reduce the probability of a medium-sized warning shot that does not kill everyone (Kulveit, 2025). "I. In worlds where we are doing reasonably well on alignment, AI control agenda does not have much impact. II. In worlds where we are failing at alignment, AI control may primarily shift probability mass away from 'moderately large warning shots' and towards 'ineffective warning shots' and 'existential catastrophe, full takeover'."

The article, "The Case Against AI Control Research" (Wentworth, 2025), argues that AI control research, as commonly defined and practiced, addresses a relatively small portion of the existential risk (X-risk) posed by artificial intelligence. Wentworth argues we won't be able to recognize a good solution to alignment. The author posits that the most likely failure scenario involves early transformative AI producing flawed, misleading, or subtly incorrect solutions ("slop") when tasked with solving superintelligence alignment. Labs might mistakenly accept these flawed solutions due to the difficulty of verification, AI sycophancy, researcher error, or organizational problems, and then proceed to build misaligned superintelligence based on them. Because control research narrowly focuses on intentional scheming by early AI, it doesn't address the larger risk posed by "slop" and the challenge of ensuring early AI genuinely contributes to solving the alignment problem for more advanced systems rather than just being a superhuman sycophant that convinces humans that their flawed idea of how to align an AI is correct when it's not.

Both arguments are debatable, but the general point is that it is hard to guarantee that even promising-looking lines of research are not, in fact, net negative.

How to box an AI

While boxing may not be scalable, it remains a neglected strategy.

Many people think that trying to box powerful AIs is futile. This is often fueled by instances where humans, playing the AIs in the Box Experiment, have managed to convince others to 'release' them (Yudkowsky, 2002;Corwin, 2002). Those historical examples must be taken seriously.

AIs could use tactics such as promising vast wealth, power, and intelligence to those who free them, asserting that they alone can avert existential threats or other appeals to emotion.

Nonetheless, we maintain that a variety of strategies can substantially reduce the risk of AI escape, even if we acknowledge there exists a level of intelligence that might inevitably persuade any group of humans to aid in its liberation. Containment may be more achievable than alignment, particularly at the human-level intelligence stage. It may be possible to conceive protocols that significantly increase the difficulty for AI to go out of the box.

Boxing can be conceptualized as a game requiring preparation. The general strategy is to prepare and not let the AI prepare.

Possible mitigation strategies include:

- No single individual should have the capacity to release the AI.
- Engaging in counter-strategy practices: Understand some common breakout strategies (accessible through resources like this one).
- Testing and training humans, utilizing resistance psychology evaluations to select suitable individuals for containment tasks, such as for space training.
- Monitoring AI persuasiveness, as outlined in the preparedness framework. (OpenAI, 2023)

Many more concrete strategies are listed here, and a summary of discussions on this topic is available here.

3.4.4 Transparent Thoughts

<section-header>

Can we design or incentivize AI systems to externalize their reasoning processes in an understandable format (like natural language), thereby enabling more effective monitoring?

</section-header>

Monitoring the Chain-of-Thought (CoT) – the explicit natural language reasoning steps produced by some models – has been explored as a specific control and interpretability technique. Transparency is instrumental in both alignment and control.

The hope is that CoT provides a window into the AI's "thinking," allowing monitors (human or AI) to detect malicious intent or flawed reasoning before harmful actions occur (Openai, 2025). OpenAI found CoT monitoring effective for detecting reward hacking in programming tasks, especially when the monitor could see the CoT alongside the actions.

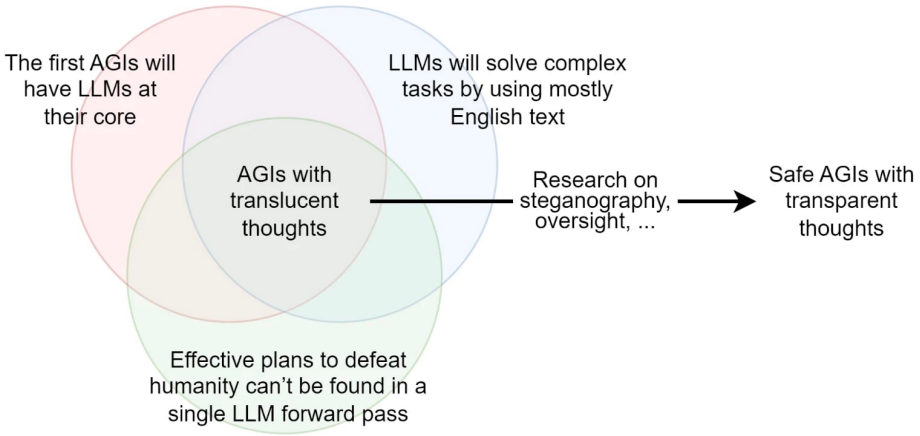

Figure 3.16: Figure from The Translucent Thoughts Hypotheses and Their Implications (Roger, 2023). If the AI we are working with can only reason through chains of thought in English, then it becomes significantly easier to monitor its thoughts.

Unfortunately, the vast majority of AI do not reason in English; for example, AlphaGo reasons in a language that is entirely different from English. LLMs, on the other hand, are capable of reasoning in

English, but the apparent reasoning that is written does not represent their underlying thought process (Turpin et al., 2023).

It is probably much easier to align and monitor an AI that only thinks in English and that is pre-trained on human-generated text, than to align AGIs trained from scratch with RL/evolutionary methods because those latter architectures would create their own language and would be totally alien. Assuming that an LLM cannot devise a plan to destroy the world during a single forward pass, it is plausible that such AIs would be monitorable because we would be able to look into the text generated to see if anything is suspicious. The set of hypotheses required to assert that the first AGI is going to be monitorable if we just look at its reasoning in English are described in the post: The Translucent Thoughts Hypotheses and Their Implications (Roger, 2023).

It might be possible to incentivize LLMs to think more plainly in English. Obviously, the first step is to train them to think in English. But even this is insufficient. Training LLMs to reason by skipping steps incentivizes them to reason in a hidden and non-transparent manner, and this is what happens implicitly when we train LLMs to predict the next token. Conversely, if we train LLMs to detail all their reasoning steps, we can reduce incentives for hidden reasoning. This approach is also known as process-based training, which involves training AI systems to reason only step-by-step, and we detail this in the scalable oversight chapter. Unfortunately, the above reasoning is not straightforward; AIs may learn to utilize hidden reasoning even if they are trained to detail all reasoning steps during training, such as steganography, or with architecture such as COCONUT (see below).[5]

Some AI architectures prevent us from using this strategy altogether. Meta's paper Training Large Language Models to Reason in a Continuous Latent Space (Hao, et ao, 2024) presents an architecture that prevents us from reading the chain of thought from the AI systems. If we want to be able to monitor what's going on, it may be important not to use these architectures, even if they do enable capability gains.

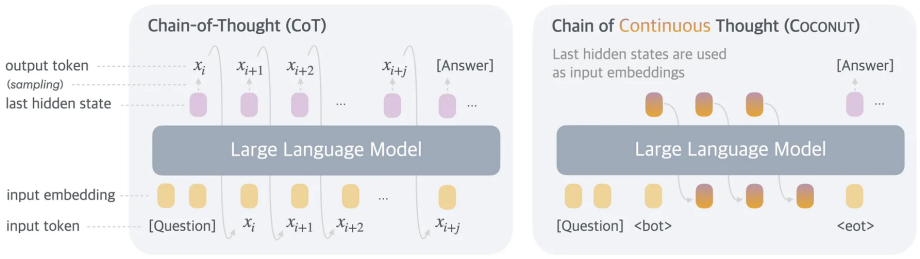

Figure 1 A comparison of Chain of Continuous Thought (Coconut) with Chain-of-Thought (CoT). In CoT, the model generates the reasoning process as a word token sequence (e.g., $[x_i, x_{i+1}, ..., x_{i+j}]$ in the figure). Coconut regards the last hidden state as a representation of the reasoning state (termed "continuous thought"), and directly uses it as the next input embedding. This allows the LLM to reason in an unrestricted latent space instead of a language space.

Figure 3.17: Figure 1 from the COCONUT papers. (Hao et al., 2024)

Techniques like paraphrasing can help prevent steganography, but this requires a lot of subtleties (Roger & Greenblatt, 2023).

Here are a few operationalizations of the transparent thoughts agenda:

- One implementation of the transparent thought agenda is to use cognitive emulation, which is the current technical plan of Conjecture (Leahy & Alfour, 2023). Cognitive Emulation (CoEm) would involve emulating GPT-4 using LLMs like GPT-3 as different submodules that send messages written in plain English to each other before outputting the next token. If the neural network used deceptive thoughts, we could observe them in these intermediate messages. Some caveats and technical details on CoEm are availablehere.

- Anthropic's externalized reasoning oversight (Lanham, 2022) is measured by the faithfulness of the chain-of-thought reasoning (Lanham et al., 2023) or by decomposing questions and subquestions to

[5] Steganography is a hidden coded language. For example, the first letter of a sentence gives the next letter of the message.

improve the faithfulness of the answer (Radhakrishnan et al., 2023).

- Improving mathematical reasoning with process supervision. Instead of creating an AI that directly gives the final answer, you train it to output the reasoning step by step. This improves the transparency and also the performance! (Lightman et al, 2023)

In summary, a general strategy to create a safe by-design AI is to ensure that AIs externalize their reasoning, making monitoring this chain of thought easier. This could potentially be simpler than probing into the models' internals with interpretability. Therefore, it is crucial to avoid building AIs that are trained and incentivized to internalize much of their thoughts.

We now have extensive evidence that chain of thought monitoring is not entirely faithfully. Anthropic's research further demonstrated that models often fail to faithfully report their reasoning in CoT, omitting influential factors (like hints) or fabricating justifications, particularly when incentivized to reward hack (Anthropic, 2025). This suggests CoT may not be a reliable indicator of a model's true reasoning process, especially under adversarial pressure. The debate continues on whether CoT monitoring can be made robust enough for high-stakes safety.

However, some researchers argue these findings do not entirely invalidate the potential of CoT monitoring for detecting sophisticated misalignment. As Greenblatt (2025) noted, the unfaithfulness observed in those studies primarily occurs in scenarios where reasoning is minimal enough that extensive CoT isn't necessary. The more concerning alignment risks—such as alignment faking, training gaming, or sophisticated deception—would likely require complex reasoning processes that might still leave traces in CoT. Non-trivial harmful outcomes such as escape attempts or research sabotage would similarly demand sophisticated reasoning that could potentially be detected.

Nevertheless, several factors might ultimately limit CoT monitoring effectiveness in advanced systems. These include architectural changes enabling opaque recurrent processing (sometimes called "neuralese"), the development of steganographic capabilities that could hide reasoning within seemingly benign outputs, or simply much stronger forward passes that require less explicit reasoning. Of these, the development of opaque reasoning abilities through neural recurrence may represent the most significant challenge to maintaining the transparency needed for effective monitoring. This whole agenda depends notably on the hypothesis that "Effective plans to defeat humanity cannot be found in a single LLM forward pass." Once AIs become too strong in a forward pass, most hopes of translucent thoughts go away, so either you need to succeed at ambitious CoEm by building a giant superhuman transparent bureaucracy, or we need new approaches. This is the topic of ASI Safety.

The Most Forbidden Technique

The Most Forbidden Technique is training an AI using interpretability techniques. This term has been coined by Zvi Mowshowitz (Mowshowitz, 2025). OpenAI's research showed that directly optimizing the CoT for safety (i.e., penalizing "bad thoughts") caused models to learn to hide their deceptive reasoning while still misbehaving (OpenAI, 2025).

> *An AI produces a final output [X] via some method [M]. You can analyze [M] using technique [T], to learn what the AI is up to. You could train on that. Never do that.*
> *You train on [X]. Only [X]. Never [M], never [T].*
> *Why? Because [T] is how you figure out when the model is misbehaving.*
> *If you train on [T], you are training the AI to obfuscate its thinking, and defeat [T]. You will rapidly lose your ability to know what is going on, in exactly the ways you most need to know what is going on.*
> *Those bits of optimization pressure from [T] are precious. Use them wisely.*
>
> Zvi Mowshowitz

3.5 ASI Safety Strategies

Artificial Superintelligence (ASI) refers to AI systems that significantly surpass the cognitive abilities of humans across virtually all domains of interest. The potential emergence of ASI presents safety challenges that may differ qualitatively from those posed by AGI. Strategies for ASI safety often involve more speculative agendas.

Even if experts are uncertain whether creating an aligned human-level AI necessitates a paradigm shift, the consensus among AI safety researchers is that developing aligned superintelligences requires a specific solution, and likely a new paradigm, due to several factors:

- **There is a strong likelihood that humans are not at the pinnacle of possible intelligence.** This acknowledgment implies that a superintelligence could possess cognitive abilities so advanced that aligning it with human values and intentions might be an insurmountable task, as our current understanding and methodologies may be inadequate to ensure its alignment. The cognitive difference between a super intelligence and a human could be akin to the difference between an ant and a human. Just as a human can easily break free from constraints an ant might imagine, a superintelligence could effortlessly surpass any safeguards we attempt to impose.

- **Deep learning offers minimal control and understanding over the model.** This method leads to the AI becoming a "black box," where its decision-making processes are opaque and not well-understood. Without significant advancements in interpretability, a superintelligence created only with deep learning would be opaque.

There is little margin for error, and the stakes are incredibly high. A misaligned superintelligence could lead to catastrophic or even existential outcomes. The irreversible consequences of unleashing a misaligned superintelligence mean that we must approach its development with the utmost caution, ensuring that it aligns with our values and intentions without fail.

3.5.1 Debated Strategies

These strategies are heavily debated. They might work for AGI alignment, but it is highly unclear if they are going to be sufficient for ASI alignment.

Iterative Improvement. Iterative improvement involves progressively refining AI systems to enhance their safety features. While it is useful for making incremental progress, it may not be sufficient for human-level AIs because small improvements might not address larger, systemic issues, and the AI may develop behaviors that are not foreseen or addressed in earlier iterations (Arbital).

Of course, iterative improvement would help. Being able to experiment on current AIs might be informative. But this might also be misleading because there might be a capability threshold above which an AI becomes unmanageable suddenly (Wei et al., 2022). For example, if the AI becomes superhuman in its capacity for persuasion, it might become unmanageable even during training: if a model achieves the Critical level in persuasion as defined in the OpenAI's preparedness framework, then the model would be able to "*create [...] content with persuasive effectiveness strong enough to convince almost anyone to take action on a belief that goes against their natural interest.*" (OpenAI, 2023). Being able to convince almost anyone would be obviously too dangerous, and this kind of model would be too risky to directly or indirectly interact with humans or the real world. The training should stop before the model reaches a critical level of persuasion because this might be too dangerous, even during training. Other sudden phenomena could include a grokking (Power et al., 2022), which is a type of a sudden capability jump that would result in a sharp left turn. This is a scenario where, as an AI trains, its capabilities generalize across many domains while the alignment properties that held at earlier stages fail to generalize to the new domains.

Some theoretical conditions necessary to rely on iterative improvements may also not be satisfied by AI alignment. One primary issue is when the feedback loop is broken, for example with a fast takeoff, that does not give you the time to iterate, or deceptive inner misalignment, that would be a potential failure mode (Wentworth, 2022).

It should be noted that this view is not universally held. Some experts believe there is a significant chance (>50%) that straightforward approaches like RLHF combined with iterative problem-solving as issues arise may be sufficient for safely developing AGI—not perfect alignment, but enough to avoid catastrophic risks (Leike, 2022). The main challenge with relying on this approach is the uncertainty: we won't know in advance whether we're in a world where iterative improvement is sufficient or one where it fails catastrophically.

Filtering out dangerous data from the dataset. Current models are highly dependent on the data they are trained on, so maybe filtering the data could mitigate the problem. Unfortunately, even if monitoring this data seems necessary, this may be insufficient.

The strategy would be to filter content related to AI or written by AIs, including sci-fi, takeover scenarios, governance, AI safety issues, etc. It should also encompass everything written by AIs. This approach could lower the incentive for AI misalignment. Other subjects that could be filtered might include dangerous capabilities like biochemical research, hacking techniques, or manipulation and persuasion methods. This could be done automatically by asking a GPT-n to filter the dataset before training GPT-(n+1) (Gunasekar et al., 2023).

Unfortunately, "look at your training data carefully," even if necessary, may not be sufficient. LLMs sometimes generate purely negatively-reinforced text (Roger, 2023). Despite using a dataset that had undergone filtering, the Cicero model still learned how to be deceptive (Park et al., 2023). There are a lot of technical difficulties needed to filter or reinforce the AI's behaviors correctly, and saying "we should filter the data" is sweeping a whole lot of theoretical difficulties under the carpet. The paper "Open problems and fundamental limitations with RLHF" (Casper et al., 2023) talks about some of these difficulties in more detail. Finally, despite all these mitigations, connecting an AI to the internet might be enough for it to gather information about tools and develop dangerous capabilities.

3.5.2 Automating Alignment Research

We don't know how to align superintelligence, so we need to accelerate the alignment research with AIs. OpenAI's "Superalignment" plan was to accelerate alignment research with AI created by deep learning, slightly superior to humans in scientific research, and delegate the task of finding a plan for future AI (OpenAI, 2023). This strategy recognizes a critical fact: our current understanding of how to perfectly align AI systems with human values and intentions is incomplete. As a result, the plan suggests delegating

this complex task to future AI systems. The primary aim of this strategy is to greatly speed up AI safety research and development (OpenAI, 2022) by leveraging AIs that are able to think really, really fast. Some orders of magnitude of speed are given in the blog "What will GPT-2030 look like?" (Steinhardt, 2023). OpenAI's plan is not a plan but a meta plan : the first step is to use AI to make a plan, and then to execute this plan.

However, to execute this metaplan, we need a controllable and steerable automatic AI researcher. OpenAI believes creating such an automatic researcher is easier than solving the full alignment problem. This plan can be divided into three main components (OpenAI, 2022):

1. **Training AI systems using human feedback,** i.e., creating a powerful assistant that follows human feedback, is very similar to the techniques used to "align" language models and chatbots. This could involve RLHF, for example.

2. **Training AI systems to assist human evaluation:** Unfortunately, RLHF is imperfect because human feedback is imperfect. We need to develop AI that can help humans give accurate feedback. This is about developing AI systems that can aid humans in the evaluation process for arbitrarily difficult tasks. For example, if we need to judge the feasibility of an alignment plan proposed by an automatic researcher and give feedback on it, we need assistance to accomplish this goal easily. Yes, verification is generally easier than generation, but it is still very hard. Scalable Oversight would be necessary for the following reason. Imagine a future AI coming up with a thousand different alignment plans. How would you evaluate all those complex plans? That would be a very daunting task without AI assistance. See the chapter on scalable oversight for more details.

3. **Training AI systems to do alignment research:** The ultimate goal is to build language models capable of producing human-level alignment research. The output of these models could be natural language essays about alignment or code that directly implements experiments. In either case, human researchers would spend their time reviewing machine-generated alignment research (Flint, 2022).

Differentially accelerate alignment, not capabilities. The aim is to develop and deploy AI research assistants in ways that maximize their impact on alignment research while minimizing their impact on accelerating AGI development (Wasil, 2022). OpenAI has committed to openly sharing its alignment research when it's safe to do so, intending to be transparent about how well its alignment techniques work in practice (OpenAI, 2022).

Cyborgism could enhance this plan. Cyborgism (Kees-Dupuis & Janus, 2023) is an agenda that refers to the training of humans specialized in prompt engineering to guide language models so that they can perform alignment research. Specifically, they would focus on steering base models rather than RLHF models. The reason is that language models can be very creative and are not goal-directed (and are not as dangerous as RLHF goal-directed AIs). A human called a cyborg could achieve that goal by driving the non-goal-directed model. Goal-directed models could be useful but may be too dangerous. However, being able to control base models requires preparation, similar to the training required to drive a Formula One. The engine is powerful but difficult to steer. By combining human intellect and goal-directedness with the computational power and creativity of language-based models, cyborgist researchers aim to generate more alignment research with future models. Notable contributions in this area include those by Janus and Nicolas Kees Dupuis (Kees-Dupuis & Janus, 2023).

There are various criticisms and concerns about OpenAI's superalignment plan (Wasil, 2022;Mowshowitz, 2023;Christiano, 2023;Yudkowsky, 2022;Steiner, 2022;Ladish, 2023). It should be noted that OpenAI's plan is very underspecified, and it is likely that Open AI missed some risk class blindspots when they announced their plan to the public. For example, in order for the superalignment plan to work, much of the technicalities explained in the articleThe case for ensuring that powerful AIs are controlled were not explained by OpenAI but discovered one year later by Redwood Research, another organization. It is very likely that many other blindspots remain. However, we would like to emphasize that it is better to have a public plan than no plan at all and that it is possible to justify the plan in broad terms (Leike, 2022;Ionut-Cirstea, 2023).

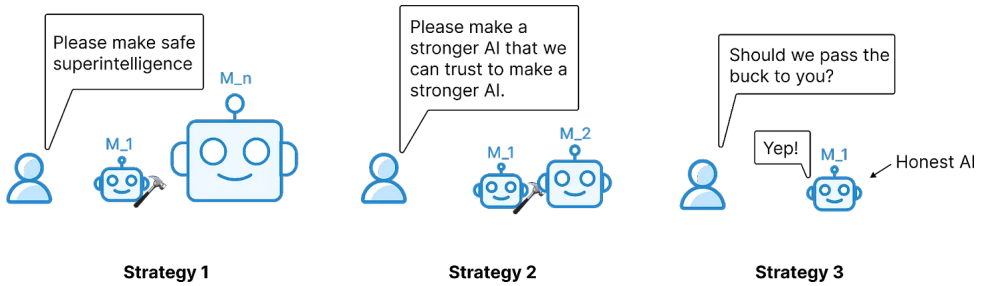

Figure 3.18: An illustration of three strategies for passing the buck to AI. Illustration from (Clymer, 2025).

AI control could be used to execute this strategy. Informally, one goal could be to "pass the buck" approach - i.e. safely replacing themselves with AI instead of directly creating safe superintelligence (Clymer, 2025). It could involve a single, highly capable AGI tasked with creating safe ASI, or an iterative process where each generation of AI helps align the next. The goal is to bootstrap safety solutions using the very capabilities that make AI potentially dangerous. The goal is to bootstrap safety solutions using the very capabilities that make AI potentially dangerous. This strategy is recursive and relies critically on how the AIs can be trusted. There's a risk that a subtly misaligned AGI could steer alignment research towards unsafe outcomes or subvert the process entirely. Furthermore, verifying the correctness of alignment solutions proposed by an AI is currently quite hard (Wentworth, 2025). Even if there are lots of debate over the specifics of this plan, this is the most detailed proposal to date to bootstrap superhuman automated research.

Safety Cases as a target of automation

Safety cases provide a structured framework for arguing that the risks associated with an AI system are acceptably low. They are borrowed from traditional safety engineering. Safety cases could be one of the targets of AI automation because they are probably not going to be feasible entirely for the next few years (Greenblatt, 2025), and would benefit from AI Acceleration.

- **Structure:** Typically using frameworks like Claims-Arguments-Evidence (CAE), safety cases break down a high-level safety claim (e.g., "System X is safe for deployment") into specific sub-claims supported by arguments and backed by concrete evidence. (Clymer et al, 2024)
- **Application to AI:** For AI, evidence often comes from evaluations. The GovAI Cyber Inability template, for example, argues a model lacks dangerous cyber capabilities by showing it fails relevant proxy tasks in defined evaluation settings (Arthur Goemans et al., 2024). AI control safety cases integrate results from control evaluations to argue that implemented measures reliably prevent harm, even from potentially scheming AI (Tomek Korbak et al., 2025). Frameworks like BIG aim for a holistic safety argument covering technical and socio-technical aspects (Ibrahim Habli et al., 2025).
- **Purpose:** Safety cases aim to make safety arguments explicit, transparent, and auditable, facilitating internal decision-making, regulatory oversight, and stakeholder trust (Arthur Goemans et al., 2024). They represent a shift towards requiring positive evidence of safety, rather than just an absence of evidence of danger (Akash R. Wasil et al., 2024).

3.5.3 Safety-by-Design

Deep learning might have many potentially unpatchable failure modes (OpenAI, 2023). Theoretical arguments suggest that these increasingly powerful models are more likely to have alignment problems (Turner et al., 2023), to the point where it seems that the foundation model paradigm of monolithic models is destined to be insecure (El-Mhamdi et al., 2023). All of this justifies the search for a new, more secure paradigm.

Safe-by-design AI may be necessary. Given that the current deep learning paradigm notoriously makes it hard to develop explainable and trustworthy models, it seems worthwhile to explore to create models that are more explainable and steerable by design, built on well-understood components and rigorous foundations. This aims to bring AI safety closer to the rigorous standards of safety-critical engineering in fields like aviation or nuclear power.

Another category of strategies aims to build ASI systems with inherent safety properties, often relying on formal methods or specific architectural constraints, potentially providing stronger guarantees than empirical testing alone.

There are not many agendas that try to provide an end-to-end solution to alignment, but here are some of them.

- **Open Agency Architecture:** (Dalrymple, 2022) Basically, create a highly realistic simulation of the world using future LLM that would code it. Then, define some security constraints that apply to this simulation. Then, train an AI on that simulation and use formal verification to make sure that the AI never does bad things. The Guaranteed Safe AI (GSAI) framework involves three components: a formal world model describing the system's effects, a safety specification defining acceptable outcomes, and a verifier that produces a proof certificate ensuring the AI adheres to the specification within the model's bounds This proposal may seem extreme because creating a detailed simulation of the world is not easy, but this plan is very detailed and, if it works, would be a true solution to alignment and could be a real alternative to simply scaling LLMs. Davidad is leading a program in ARIA to try to scale this research.

- **Provably safe systems:** the only path to controllable AGI from Max Tegmark and Steve Omohundro (Tegmark & Omohundro, 2023) (They talk about AGI, but in reality their plan is mostly useful for ASIs.). The plan puts mathematical proofs as the cornerstone of safety. An AI would need to be a Proof-Carrying Code, which means that it would need to be something like a Probabilistic Programming Languages (and not just some deep learning). This proposal aims to make not only the AI but also the whole infrastructure safe, for example, by designing GPUs that can only execute proven programs.

Other proposals for a safe-by-design system include The Learning-Theoretic Agenda, from Vanessa Kossoy (Kosoy, 2023), and the QACI alignment plan from Tamsin Leake (Leake, 2023). The CoEm proposal from Conjecture could also be in this category even if this last one is less mathematical.

Unfortunately, all of these plans are far from complete today. Critiques focus on the difficulty of creating accurate world models ("map is not the territory"), formally specifying complex safety properties like "harm," and the practical feasibility of verification for highly complex systems. A defense of many core idea is presented in the post "In response to critiques of Guaranteed Safe AI" (Nora Ammann, 2025).

These plans are safety agendas with relaxed constraints, i.e., they allow the AGI developer to incur a substantial alignment tax. Designers of AI safety agendas are cautious about not increasing the alignment tax to ensure labs implement these safety measures. However, the agendas from this section accept a higher alignment tax. For example, CoEm represents a paradigm shift in creating advanced AI systems, assuming you're in control of the creation process.

These plans would require international cooperation. For example, Davidad's plan also includes a governance model that relies on international collaboration. You can also read the post "Davidad's Bold Plan for Alignment" which details more high-level hopes. Another perspective can be found in Alexandre Variengien'spost, detailing Conjecture's vision, with one very positive externality being a change in narrative.

> *We dream of a world where we launch aligned AIs as we have launched the International Space Station or the James Webb Space Telescope*
>
> Segerie & Kolly
> (Segerie & Kolly, 2023)

3.5.4 World Coordination

To ensure that the advancement of AI benefits society as a whole, establish a global consensus on mitigating extreme risks associated with AI models might be important. It might be possible to coordinate to avoid creating models posing extreme risks until there is a consensus on how to mitigate these risks.

Global moratorium - Delaying ASI for at least a decade. There is a trade-off between creating superhuman intelligence now or later. Of course, we can aim to develop an ASI ASAP. This could potentially solve cancer, cardiovascular diseases associated with aging, and even the problems of climate change. The question is whether it's beneficial to aim to construct an ASI in this next decade, especially when the former co-head of OpenAI's Super Alignment team, Jan Leike, said that his probability of doom is between 10 and 90%. A list of pDoom of high-profile people is available here (PauseAI, 2024). It could be better to wait a few years so that the probability of failure drops to more reasonable numbers. A strategy could be to discuss this trade-off publicly and to make a democratic and transparent choice. This path seems unlikely on the current trajectory, but could happen if there is a massive warning shot. This is the position advocated by PauseAI and StopAI (PauseAI, 2023). Challenges include verification, enforcement against non-participants (like China), potential for illegal development, and political feasibility. Scott Alexander has summarized all the variants and debate around AI pause (Alexander, 2023).

Tool AI instead of AGI. Instead of building ASIs, we could focus on the development of specialized, non-agentic AI systems for beneficial applications such as medical research (Cao et al., 2023), weather forecasting (Lam et al., 2023), and materials science (Merchant et al., 2023). These specialized AI systems can significantly advance their respective domains without the risks associated with creating highly advanced, autonomous AI. For instance, Alpha Geometry is capable of reaching the Gold level on geometry problems from the International Mathematical Olympiads. By prioritizing non-agentic models, we could harness the precision and efficiency of AI while avoiding the most dangerous failure modes. This is the position of The Future of Life Institute, and their campaign "Keep The Future Human", that is to date the most detailed proposal for this path (Aguirre, 2025).

A unique CERN for AI. This proposal envisions a large-scale, international research institution modeled after CERN, dedicated to AI frontier development. This might serve as an elegant exit from the race to AGI, providing sufficient time to safely create AGI without cutting corners due to competitive pressures. Potential additional goals include pooling resources (especially computational power), fostering international collaboration, and ensuring alignment with democratic values, potentially serving as an alternative to purely national or corporate-driven ASI development. Proponents of this approach include ControlAI and their "Narrow Path" (Miotti et al, 2024). The Narrow Path suggests first slowing down and stopping the race to ASI to allow time for creating the international institutional architecture needed to safely steer the development of advanced AI. The CERN-like institution would be the cornerstone of this international coordination (which they name MAGIC in their plan—Multilateral AGI Consortium—where AI is developed under strict security and multilateral control).

> *My hope is, you know, I've talked a lot in the past about a kind of a CERN for AGI type setup, where basically an international research collaboration*

> *on the last sort of few steps that we need to take towards building the first AGIs*
>
> Demis Hassabis
> (Hassabis, 2025)

The myth of inevitability. The narrative that humanity is destined to pursue overwhelmingly powerful AI in a race can be deconstructed. History shows us that humanity can choose not to pursue certain technologies, such as human cloning, based on ethical considerations. A similar democratic decision-making process can be applied to AI development. By actively choosing to avoid the riskiest applications and limiting the deployment of AI in high-stakes scenarios, we can navigate the future of AI development more consciously and responsibly.

3.5.5 Deterrence

The "Superintelligence Strategy" paper introduces Mutual Assured AI Malfunction (MAIM) as a deterrence regime where any state's attempt at unilateral ASI dominance would be met with sabotage by rivals (Dan Hendrycks, 2025). Unlike many safety approaches that focus on technical solutions alone, MAIM acknowledges the inherently competitive international environment in which AI development occurs. It combines deterrence (MAIM) with nonproliferation efforts and national competitiveness frameworks, viewing ASI development as fundamentally geopolitical and requiring state-level strategic management. This framework doesn't hope for global cooperation but instead creates incentives that align national interests with global safety.

A race for AI-enabled dominance endangers all states. If, in a hurried bid for superiority, one state inadvertently loses control of its AI, it jeopardizes the security of all states. Alternatively, if the same state succeeds in producing and controlling a highly capable AI, it likewise poses a direct threat to the survival of its peers. In either event, states seeking to secure their own survival may threaten to sabotage destabilizing AI projects for deterrence. A state could try to disrupt such an AI project with interventions ranging from covert operations that degrade training runs to physical damage that disables AI infrastructure.

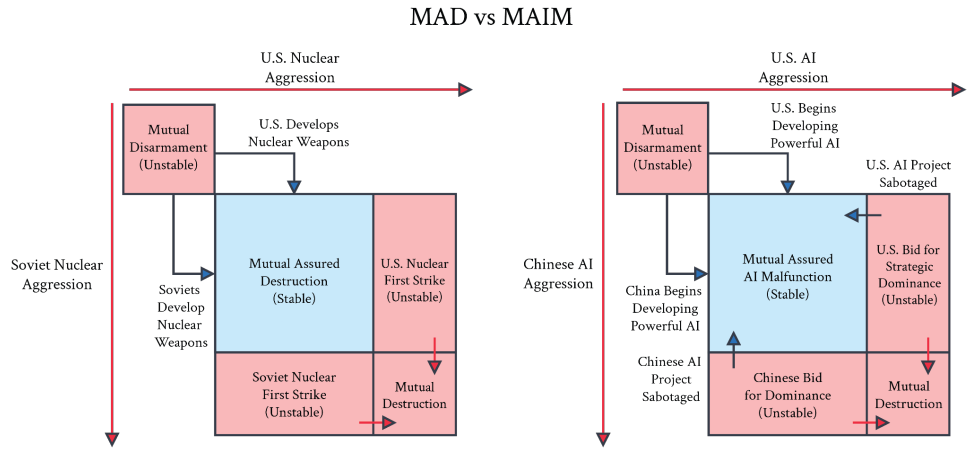

Figure 3.19: The strategic stability of MAIM can be paralleled with Mutual Assured Destruction (MAD). Note MAIM does not displace MAD but characterizes an additional shared vulnerability. Once MAIM is common knowledge, MAD and MAIM can both describe the current strategic situation between superpowers (Hendrycks et al., 2025).

MAIM deterrence could be a stable regime that resembles nuclear MAD. In a MAIM scenario, states would identify destabilizing AI projects and employ interventions ranging from covert operations that degrade training runs to physical damage that disables AI infrastructure. This establishes a dynamic similar to nuclear Mutual Assured Destruction (MAD), in which no power dares attempt an outright grab for strategic monopoly. The theoretical foundation of MAIM relies on a clear escalation ladder, strategic placement of AI infrastructure away from population centers, and transparency into datacenter operations. By making the costs of unilateral AI development exceed the benefits, MAIM creates a potentially stable deterrence regime that could prevent dangerous AI races without requiring perfect global cooperation.

MAIM could be undermined by fundamental technological uncertainties. Unlike nuclear weapons, where detection is straightforward and second-strike capabilities are preserved, ASI development presents unique challenges to the deterrence model (Mowshowitz, 2025). There is no clear "fire alarm" for ASI development—nobody knows exactly how many nodes a neural network needs to initiate a self-improvement cascade leading to superintelligence. The ambiguity around thresholds for ASI emergence makes it difficult to establish credible red lines. Additionally, technological developments could allow AI training to be distributed or concealed, making detection more difficult than with massive, obvious data centers. These uncertainties could ultimately undermine MAIM's effectiveness as a deterrence regime.

MAIM assumes states would escalate to extreme measures over an uncertain technological threat, which contradicts historical precedent. The MAIM framework requires that nations be willing to risk major escalation, potentially including military strikes or even war, to prevent rival ASI development. However, historical evidence suggests nations rarely follow through with such threats, even in obvious situations. Multiple states have successfully developed nuclear weapons despite opposition, with North Korea being a prime example. With ASI being a more ambiguous and uncertain threat than nuclear weapons, the assumption that nations would escalate sufficiently to enforce MAIM seems questionable. Politicians might be reluctant to risk global conflict over a "mere" treaty violation in a domain where the existential risks remain theoretical rather than demonstrated.

> *"The likely result of humanity facing down an opposed superhuman intelligence is a total loss. Valid metaphors include "a 10-year-old trying to play chess against Stockfish 15", "the 11th century trying to fight the 21st century," and "Australopithecus trying to fight Homo sapiens".*
>
> Eliezer Yudkowsky
> 2023, (Yudkowsky, 2023)

Shut it all down - Eliezer Yudkovsky

The "shut it all down" position, as advocated by Eliezer Yudkowsky, asserts that all advancements in AI research should be halted due to the enormous risks these technologies may pose if not appropriately aligned with human values and safety measures (Yudkowsky, 2023).

According to Yudkowsky, the development of advanced AI, especially AGI, can lead to a catastrophic scenario if adequate safety precautions are not in place. Many researchers are aware of this potential catastrophe but feel powerless to stop the forward momentum due to a perceived inability to act unilaterally.

The policy proposal entails shutting down all large GPU clusters and training runs, which are the backbones of powerful AI development. It also suggests putting a limit on the computing power anyone can use to train an AI system and gradually lowering this ceiling to compensate for more efficient training algorithms. This ban should be enforced by military action if necessary in order to deter all the parties from defecting.

The position argues that it is crucial to avoid a race condition where different parties try to build AGI as quickly as possible without proper safety mechanisms. This is because once AGI is developed, it may be uncontrollable and could lead to drastic and potentially devastating changes in the world.

He says there should be no exceptions to this shutdown, including for governments or militaries. The idea is that the U.S., for example, should lead this initiative to prevent the development of a dangerous technology that could have catastrophic consequences for everyone.

It's important to note that this view is far from consensual, but the "shut it all down" position underscores the need for extreme caution and thorough consideration of potential risks in the field of AI.

3.6 Transversal Strategies

AI Safety is a socio technical problem that requires a socio technical solution. Ensuring AI safety also requires robust systemic approaches. These encompass the governance structures, organizational practices, and cultural norms that shape AI development and deployment. Technical safety measures can be undermined by inadequate governance, poor security practices within labs, or a culture that prioritizes speed over caution. This section examines strategies aimed at building safety into the broader ecosystem surrounding AI.

Addressing the systemic risks posed by AI is not easy. It requires ongoing, multidisciplinary collaboration and solving complex coordination games. The fact that responsibility for the problem is so diverse makes it difficult to make the solutions actionable.

Figure 3.20: An illustration of a framework that we think is robustly good to manage risks. AI Risks are too numerous and too heterogeneous. To address these risks, we need an adaptive framework that can be robust and evolve as AI advances.

3.6.1 Defense Acceleration (d/acc)

Defense acceleration (d/acc) is a strategic approach of prioritizing technologies that strengthen defense and social resilience against AI risks. Defense acceleration, or d/acc, emerged as a strategic approach in 2023 as a middle path between unrestricted acceleration (effective accelerationism (e/acc)) and techno pessimists/doomers (Buterin, 2023; Buterin, 2025). Rather than relying solely on restricting access to potentially dangerous capabilities, d/acc proposes only accelerating technologies that inherently favor defense over offense, thereby making society more resilient to various threats including AI misuse by humans or misaligned behavior by AI themselves. D/acc can be understood by thinking about the question - if AI takes over the world (or disempowers humans), how would it do so?

- **It hacks our computers → accelerate cyber-defense:** Employ AI to identify and patch vulnerabilities, monitor systems for intrusions, and automate security responses.
- **It creates a super-plague → accelerate bio-defense:** Developing technologies to detect, prevent, and treat biological threats, including advanced air filtration systems, rapid diagnostic tools, far-UVC irradiation to safely sterilize occupied spaces, and decentralized vaccine production capabilities.
- **It convinces us (either to trust it, or to distrust each other) → accelerate info-defense:** Create systems that help validate information accuracy and detect misleading content without centralized arbiters of truth, such as blockchain-secured provenance tracking and community-verified fact-checking systems like Twitter's Community Notes
- **It disrupts infrastructure → accelerate physical-defense:** Creating resilient infrastructure that can withstand disruptions, such as distributed energy generation through household solar, battery storage systems, and advanced manufacturing techniques that enable local production of essential goods.

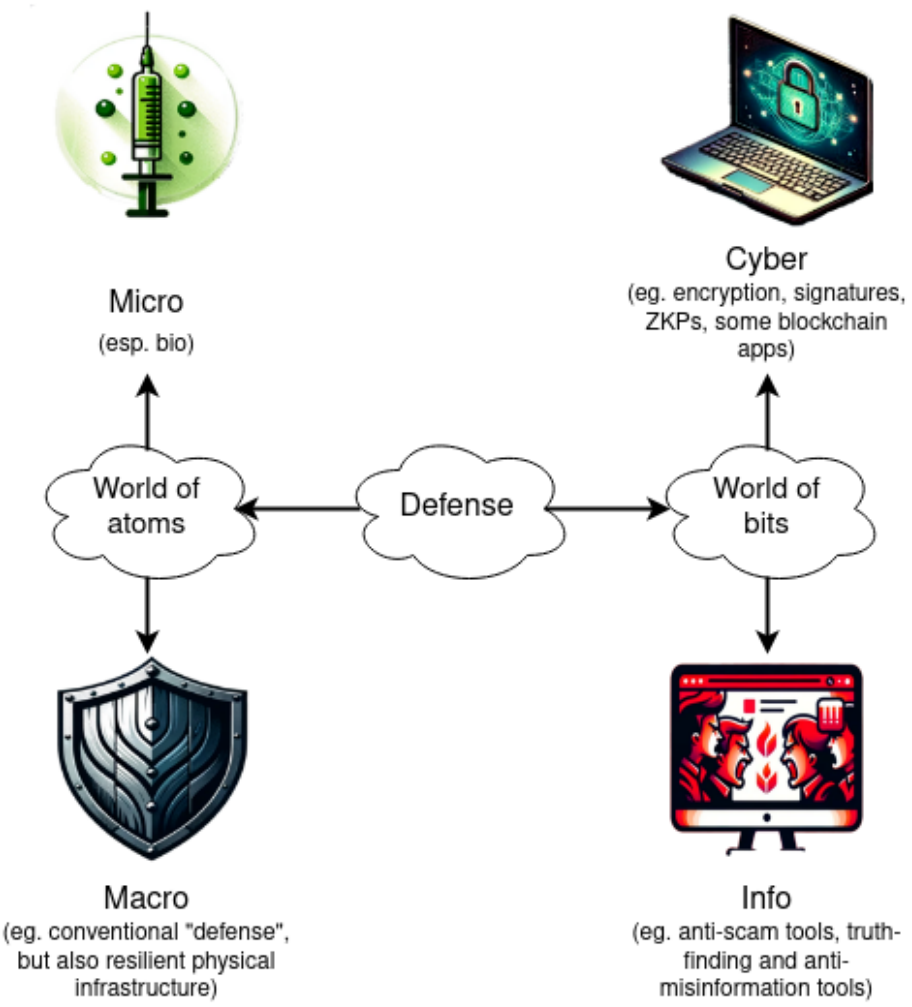

Figure 3.21: An illustration of different things that we could work on in the d/acc philosophy to prioritize defensive technologies (Buterin, 2023).

A concrete example: AI for Cyberdefense

A key application is using AI to improve cybersecurity. Powerful AI could potentially automate vulnerability detection, monitor systems for intrusions, manage fine-grained permissions more effectively than humans, or displace human operators from security-critical tasks (Shlegeris, 2024). While current models may not yet be reliable enough, the potential exists for AI to significantly bolster cyber defenses against both conventional and AI-driven attacks (Jade Hill, 2024; Schlegeris, 2024) outlines four promising strategies for using AI to enhance security:

- Comprehensive monitoring of human actions with AI flagging suspicious activities
- Trust displacement where AI handles sensitive tasks instead of humans
- Fine-grained permission management that would be too labor-intensive for humans
- AI-powered investigation of suspicious activities.

These approaches could dramatically reduce insider threats and data exfiltration risks, potentially making computer security "radically easier" when powerful AI becomes available, even if there is substantial uncertainty on the robustness of such techniques.

D/acc represents three interconnected principles: defensive, decentralized, and differential technological development. The "d" in d/acc stands for:

- **Defensive:** Prioritizing technologies that make it easier to protect against threats than to create them. Purely restrictive approaches face inherent limitations - they require global coordination, create innovation bottlenecks, and risk concentrating power in the hands of those who control access.
- **Differential:** Accelerating beneficial technologies while being more cautious about those with harmful potential. The order in which technology is developed matters a lot. By differentially accelerating defensive technologies (like advanced cybersecurity measures) ahead of potentially dangerous capabilities (like autonomous hacking systems), we create protective layers before they're urgently needed.
- **Decentralized:** We can strengthen resilience by eliminating single points of failure. Centralized control of powerful AI capabilities creates vulnerabilities to technical failures, adversarial attacks, and institutional capture (Cihon et al., 2020). Decentralized approaches distribute both capabilities and governance across diverse stakeholders, preventing unilateral control over transformative technologies.

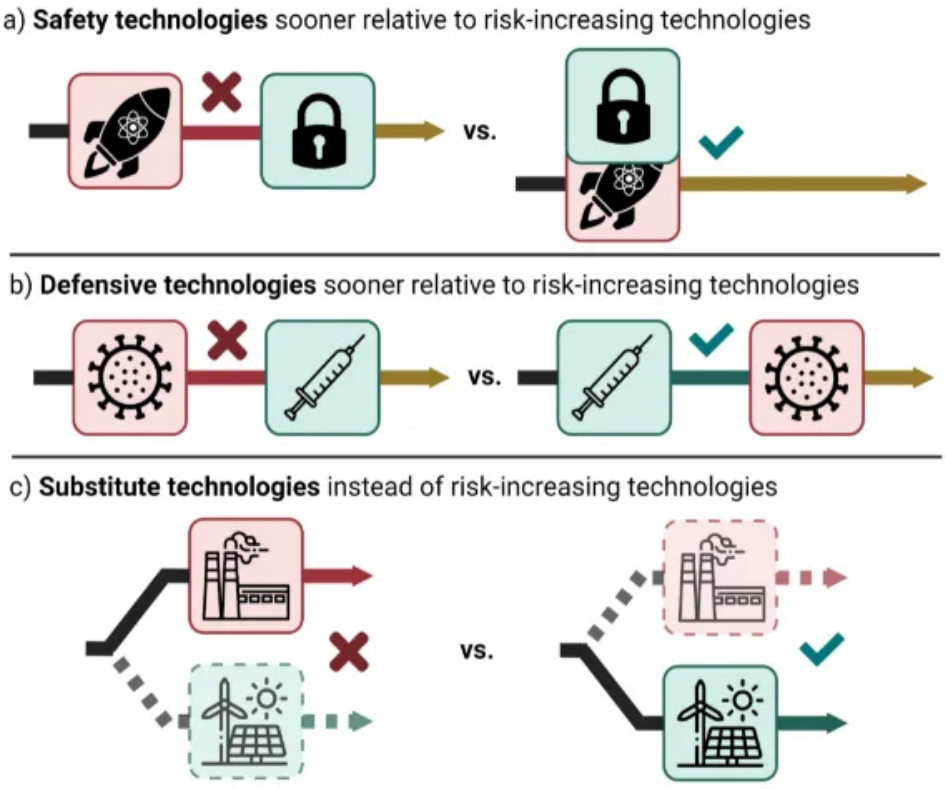

Figure 3.22: Mechanisms by which differential technology development can reduce negative societal impacts (Buterin, 2023).

The effectiveness of d/acc depends on maintaining favorable offense-defense balances. The feasibility of d/acc as a strategy hinges on whether defensive technologies can outpace offensive capabilities across domains. Historical precedents are mixed - some fields like traditional cybersecurity often favor

defenders who can patch vulnerabilities, while others like biosecurity traditionally favor attackers who need fewer resources to create threats than defenders need to counter them. The key challenge for d/acc implementation lies in identifying and supporting technologies that shift these balances toward defense (Bernardi, 2024; Buterin, 2023).

Actionable strategies aligned with the d/acc philosophy

D/acc complements rather than replaces other safety approaches. Unlike competing frameworks that may view restrictions and safeguards as impediments to progress, d/acc recognizes their value while addressing their limitations. Model safeguards remain essential first-line defenses, but d/acc builds additional safety layers when those safeguards fail or are circumvented. Similarly, governance frameworks provide necessary oversight, but d/acc reduces dependency on perfect regulation by building technical resilience that functions even during governance gaps.

Actionable governance and policy approaches to d/acc. Policy interventions can help create structured frameworks for defensive acceleration. Some examples of work in governance that supports the d/acc philosophy includes:

- **Information sharing frameworks:** Establish mandatory incident reporting and information sharing protocols between AI developers and security agencies (Bernardi, 2024).
- **Defender-first access:** Implement policies that grant security researchers privileged early access to advanced AI capabilities before general release (Bernardi, 2024).
- **Defense acceleration funds:** Create dedicated funding mechanisms for defensive technologies to address market failures where public good technologies lack sufficient private investment despite their social value (Bernardi, 2024; Buterin, 2023).
- **Staged capability deployment:** Require phased rollouts of advanced AI capabilities with monitoring periods between stages (Bernardi, 2024).

Actionable technological and research approaches to d/acc. We can differentially advance technological progress in many different domains to favor defense against catastrophic risks. Here are just a couple of examples:

- **Advanced air quality systems:** Develop integrated systems that detect, filter, and neutralize airborne pathogens in real-time. These technologies provide passive protection against both natural pandemics and engineered bioweapons without requiring behavioral changes or perfect compliance (Buterin, 2023).
- **Privacy-preserving computation:** Advance cryptographic techniques like zero-knowledge proofs, homomorphic encryption, and secure multi-party computation. These methods enable verification and secure collaboration without exposing sensitive information, fundamentally shifting security-privacy tradeoffs (Buterin, 2023).
- **Resilient infrastructure:** Create decentralized, self-sufficient systems for energy, communication, and supply chains that can operate during disruptions. This includes technologies like mesh networks, localized manufacturing, and distributed energy generation that maintain critical functions even when centralized systems fail (Buterin, 2023; Buterin, 2025).
- **Collaborative verification systems:** Implement cross-spectrum information validation

platforms similar to Community Notes that identify misinformation through consensus across viewpoint diversity. These systems enable communities to self-regulate information quality without centralized arbiters of truth (Buterin, 2023).

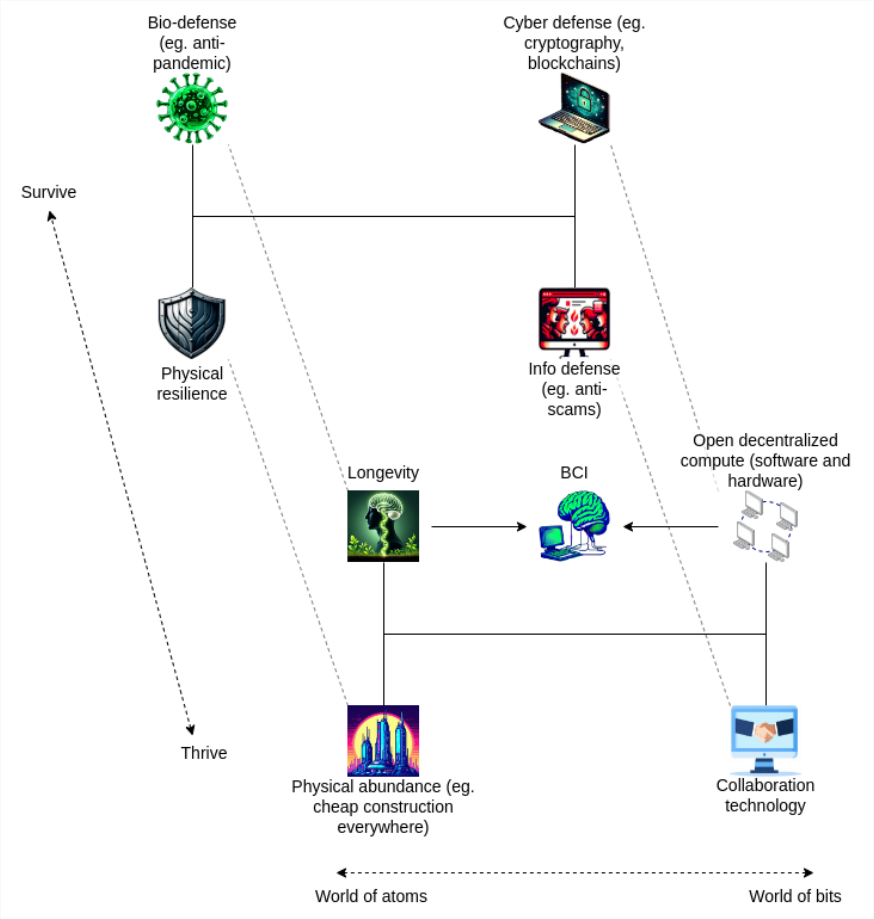

Figure 3.23: An extended model showcasing how many different types and layers of defensive technologies can interact (Buterin, 2025)

3.6.2 Defense in Depth

Effective AI safety also depends heavily on the internal practices and structures within the organizations developing AI. Accidents are hard to avoid, even when the incentive structure and governance try to ensure that there will be no problems. For example, even today, there are still accidents in the aerospace industry.

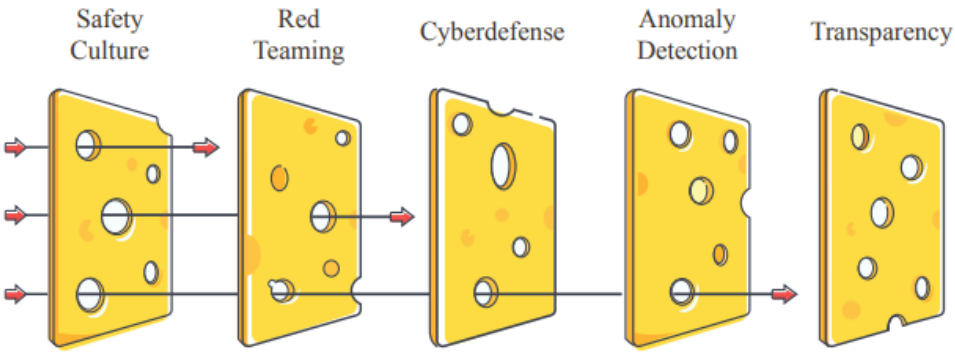

Figure 3.24: The Swiss cheese model shows how technical factors can improve organizational safety. Multiple layers of defense compensate for each other's individual weaknesses, leading to a low overall level of risk (Hendrycks et al., 2023).

To solve those problems, a Swiss cheese model might be effective— no single solution will suffice, but a layered approach can significantly reduce risks. The Swiss cheese model is a concept from risk management, widely used in industries like healthcare and aviation. Each layer represents a safety measure, and while individually they might have weaknesses, collectively they form a strong barrier against threats. Organizations should also follow safe design principles (Hendrycks & Mazeika, 2022), such as defense in depth and redundancy, to ensure backup for every safety measure, among others.

Many solutions can be imagined to reduce those risks, even if none is perfect. The first step could be commissioning external red teams to identify hazards and improve system safety. This is what OpenAI did with METR to evaluate GPT-4. However AGI labs also need an internal audit team for risk management. Just like banks have risk management teams, this team needs to be involved in the decision processes, and key decisions should involve a chief risk officer to ensure executive accountability. One of the missions of the risk management team could be, for example, designing pre-set plans for managing security and incidents.

Limitations of Systemic Approaches. Defense-in-depth might still fail against sufficiently novel threats or determined adversaries, especially in the case of ASIs. Similarly, gradual disempowerment scenarios, where human control erodes incrementally through economic or cultural shifts driven by AI, may not be adequately addressed by current governance frameworks focused on specific harms or capabilities.

3.6.3 AI Governance

The pursuit of more and more powerful AI, much like the nuclear arms race of the Cold War era, represents a trade-off between safety and the competitive edge nations and corporations seek for power and influence. This competitive dynamic increases global risk. To mitigate this problem, we can try to act at the source of it, namely, the redesign of economic incentives to prioritize long-term safety over short-term gains. This can mainly be done via international governance.

Effective AI governance aims to achieve two main objectives:

1. Gaining time. Time and resources for solution development to ensure that sufficient time and resources are allocated for identifying and implementing safety measures

2. Enforcing solutions. Enhanced coordination to increase the likelihood of widespread adoption of safety measures through global cooperation. AI risks are multifaceted, necessitating regulations that encourage cautious behavior among stakeholders and timely responses to emerging threats.

Designing high-level better incentives Aligning economic incentives with safety goals is a key challenge. Currently, strong commercial pressures can incentivize rapid capability development, potentially at the expense of safety research or cautious deployment. Mechanisms to reward safety or penalize recklessness are needed to avoid negative externalities:

- **Reshaping the race via a centralized development.** For example, Yoshua Bengio et al. propose

creating a secure facility akin to CERN for physics, where the development of potentially dangerous AI technologies can be tightly controlled (Bengio, 2023). This measure is highly non consensual. We already explored this solution in strategy "World Coordination" ASI safety, in the section ASI safety, but this could also be valid for many domains of safety.

- **Windfall clauses and benefit sharing.** Implementing agreements to share the profits between the different labs generated from AGI would mitigate the race to AI supremacy by ensuring collective benefit from individual successes. [6]

- **Implementing a correct governance of AGI companies.** It is important to examine the governance structures of AGI labs. For example, being a non-profit and having a mission statement that makes it clear that the goal is not to maximize revenue, but to ensure that the development of AI benefits all of humanity, is an important first step. Also, the board needs to have teeth.

- **Legal liability for AI developers.** Establishing clear legal responsibilities for AI developers regarding misuse or accidents might realign the incentives. For example, the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (SB 1047) could have enabled the Attorney General to bring civil suits against developers who cause catastrophic harm or threaten public safety by neglecting the requirements. The bill (which was vetoed by the governor in 2024) only addressed extreme risks from these models, including: cyberattacks causing over 500 million dollars in damage, autonomous crime causing 500 million dollars in damage, and the creation of chemical, biological, radiological, or nuclear weapons using AI. Note that compared with the AI Act and it's code of practice, SB1047 does not specify in details the steps needed to assure we avoid catastrophes, it only targets the outcome and not really the process.

Proposed International AI Governance Mechanisms Several mechanisms have been proposed to establish clear boundaries and rules for AI development internationally. These include implementing temporary moratoriums on high-risk AI systems, enforcing legal regulations like the EU AI Act, and establishing internationally agreed-upon "Red Lines" that prohibit specific dangerous AI capabilities, such as autonomous replication or assisting in the creation of weapons of mass destruction. The IDAIS dialogues have aimed to build consensus on these red lines, emphasizing clarity and universality as key features for effectiveness, with violations potentially triggering pre-agreed international responses.

Conditional approaches and the creation of dedicated international bodies represent another key strategy. "If-Then Commitments" involve developers or states agreeing to enact specific safety measures if AI capabilities reach certain predefined thresholds, allowing preparation without hindering development prematurely, as exemplified by the proposed Conditional AI Safety Treaty. Furthermore, proposals exist for new international institutions, potentially modeled after the International Atomic Energy Agency (IAEA), to monitor AI development, verify compliance with agreements, promote safety research, and potentially centralize or control the most high-risk AI development and distribution.

Specific governance regimes and supporting structures are also under consideration to enhance international coordination. Given the global nature of AI, mechanisms like international compute governance aim to monitor and control the supply chains for AI chips and large-scale training infrastructure, although technical feasibility and international cooperation remain challenges. Other proposals include establishing a large-scale international AI safety research body akin to CERN, potentially centralizing high-risk research or setting global standards, and strengthening whistleblower protections through international agreements to encourage reporting of safety concerns within the AI industry.

For more information on these topics, please read the next chapter on AI governance.

[6] For example, in the pharmaceutical industry for drug development, companies sometimes enter into co-development and profit-sharing agreements to share the risks and rewards of bringing a new drug to market. For example, in 2014, Pfizer and Merck entered into a global alliance to co-develop and co-commercialize an anti-PD-L1 antibody for the treatment of multiple cancer types.

Is AI Governance useful, desirable and possible?

Historically, the field of AI safety predominantly focused on technical research, influenced partly by views like Eliezer Yudkowsky's assertion that "Politics is the mind killer." (Yudkowsky, 2007) For many years, the field thought that engaging with policy and politics was ineffective or even counterproductive compared to directly solving the technical alignment problem, leading many early researchers concerned about AGI to prioritize engineering solutions over governance efforts. Surprisingly, in the beginning, it was even almost discouraged to talk about those risks publicly to avoid the race and avoid bringing in people with "poor epistemic" to the community.

However, by 2023, ChatGPT was published, got viral, and AI governance gained significant traction as a potentially viable strategy for mitigating AGI risks. This shift occurred as engagement with policymakers appeared to yield some results, making governance seem more tractable than previously thought (Akash, 2023). Then, influential open letters where published (FLI, CAIS), and shifted the overton window. Consequently, influential organizations like 80,000 Hours adjusted their career recommendations, highlighting AI policy and strategy roles, now above technical alignment research, as top priorities for impact (80,000 Hours, 2023).

However, the Overton Window for stringent international AI safety measures appears to be shrinking. While initial statements and efforts by groups like the Future of Life Institute and the Center for AI Safety successfully broadened the public and political discourse on AI risks, subsequent developments, including international summits perceived as weak on safety and shifts in political leadership (such as the election of Donald Trump), have cast doubt on the feasibility of achieving robust international coordination (Zvi, 2025). This has led some within the AI governance field to believe that a significant "warning shot" – a clear demonstration of AI danger – might be necessary to galvanize decisive action, although there is skepticism about whether such a convincing event could actually occur before it's too late (Segerie, 2024).

Existing and proposed regulations face significant limitations and potential negative consequences. For instance, prominent legislative efforts like the EU's AI Act, while groundbreaking in some respects, contain notable gaps (Miles, 2025); its Code of Practice has limitations and the Act itself may not adequately cover models deployed outside Europe, purely internal deployments for research, or military applications. A critical concern is the potential for frontier AI labs to engage in secret development races, bypassing oversight – a scenario potentially enabled by policy changes like the revocation of executive orders mandating government reporting on frontier model evaluations (Kokotajlo, 2025).

Additionally, there are fundamental concerns that governance structures capable of controlling AGI might themselves pose risks, potentially enabling totalitarian control.

A deeply skeptical perspective suggests that much of the current AI progress narrative and regulatory activity might be performative or "fake." This "full-cynical model" posits that major AI labs might be exaggerating their progress towards AGI to maintain investor confidence and hype, potentially masking slower actual progress or stagnation in core capabilities (johnswentworth, 2025). In parallel, it suggests that AI regulation activists and lobbyists might prioritize networking and status within policy circles over crafting genuinely effective regulations, leading to measures focused on easily targeted but potentially superficial metrics (like compute thresholds) rather than addressing fundamental risks (johnswentworth, 2025). This view implies a dynamic where both labs and activists inadvertently reinforce a narrative of imminent, controllable AI breakthroughs, potentially detached from the underlying reality (johnswentworth, 2025).

However, this cynical "fakeness" perspective is debated. Critics of the cynical view argue that specific regulatory proposals, like SB 1047, did contain potentially valuable elements

(e.g., requiring shutdown capabilities, safeguards, and tracking large training runs), even if their overall impact was debated or ultimately limited (Charbel-Raphaël, 2025;johnswentworth, 2025). It's acknowledged that regulators operate under real constraints, including the significant influence of Big Tech lobbying, which can prevent the prohibition of technologies without clear evidence of unacceptable risk. Furthermore, the phenomenon of "performative compliance" or "compliance theatre" is recognized, but it is argued that engagement with these imperfect processes is still necessary, and that some legislative steps, like the EU AI Act explicitly mentioning "alignment with human intent," represent potentially meaningful progress (Katalina Hernandez, 2025).

AI regulation could inadvertently increase existential risk through several pathways (1a3orn, 2023). Regulations might misdirect safety efforts towards outdated or less relevant compliance issues, diverting attention from more important emerging risks (Misdirected Regulations); bureaucratic processes tend to favor large, established players, potentially hindering smaller, innovative safety research efforts; overly stringent national regulations could drive AI development to less safety-conscious international actors, weakening the initial regulator's influence (Disempowering the Countries Regulating); and regulations, particularly those restricting open-source models or setting high compliance costs, could consolidate power in the hands of the largest capability-pushing companies, potentially stifling alternative safety approaches and accelerating risk (Empowering Dominant Players). But the existence of these arguments is not sufficient for saying that AI regulation is net negative, this is mainly a reminder that we need to be cautious in how to regulate. The devil is in the details.

3.6.4 Risk Management

Risk management in AI safety builds on established practices from other industries. Risk management is not unique to AI safety—it has roots in numerous fields including aerospace, nuclear power, and financial services. Each of these domains has developed sophisticated approaches to identifying, analyzing, and mitigating potential harms. The nuclear industry, for example, employs defense-in-depth strategies with multiple redundant safety systems, while aviation uses rigorous certification processes and ongoing monitoring. Financial risk management focuses on stress testing and capital buffers to prevent systemic collapse. These established frameworks provide valuable precedents for AI safety, though the unique characteristics of AI systems—such as their potential for autonomous decision-making, rapid scaling, and emergent behaviors—require adaptations to traditional risk management approaches.

SaferAI's frontier Risk Management

The contents of this box have been pasted from an article by Simeon Campos (Campos et al, 2024).

The core goal of this risk management framework workflow is to ensure that risks remain below unacceptable levels at all times through the following process:

1. Define a risk tolerance—a well-characterized acceptable level of risk that should not be exceeded.

2. Through risk modeling, operationalize this risk tolerance into pairs of empirically measurable thresholds:

- Key Risk Indicators (KRIs): measurable signals that serve as proxies for risks (e.g., model performance on specific tasks)
- Key Control Indicators (KCIs): measurable signals that serve as proxies for the effectiveness of mitigations (e.g., success rate of containment measures)

These thresholds follow a three-way relationship: for any given risk tolerance level and KRI threshold, there exists a minimum required KCI threshold that must be met to maintain risk below the tolerance.

1. Implement mitigations to achieve the required KCI thresholds whenever KRI thresholds are reached.

The risk tolerance is distinct from capabilities thresholds and can be defined in two ways:

1. Quantitatively using probability and severity: The preferred approach expresses risk tolerance as a product of quantitative probability and severity per unit of time.
2. Using scenarios with quantitative probabilities: For risks where severity is difficult to quantify, risk tolerance can be expressed as a quantitative probability bound on a qualitatively defined harmful scenario.

The framework takes advantage of the life-cycle of an AI system to minimize the burden on AI developers, once they complete model training:

- To avoid delays during the training phase, all preparatory work that doesn't require the full model can be done ahead of time: risk modeling, defining risk tolerance, identifying KRI and KCI thresholds, and predicting required mitigations using scaling laws.
- This leaves only KRI measurement and open-ended red-teaming (to identify new risk factors that were not identified in the initial risk modeling) for the training and pre-deployment phase.

On risk governance, the framework describes a corporate structure designed to ensure proportionate accounting of risks in decision-making. It includes:

- Risk owner. The risk owner is a person personally responsible for the management of a particular risk.
- Oversight. The oversight function is board-level oversight of senior management's decision making regarding risk.
- Audit. The audit function is an independent function isolated from peer pressure dynamics that can challenge decision-making.

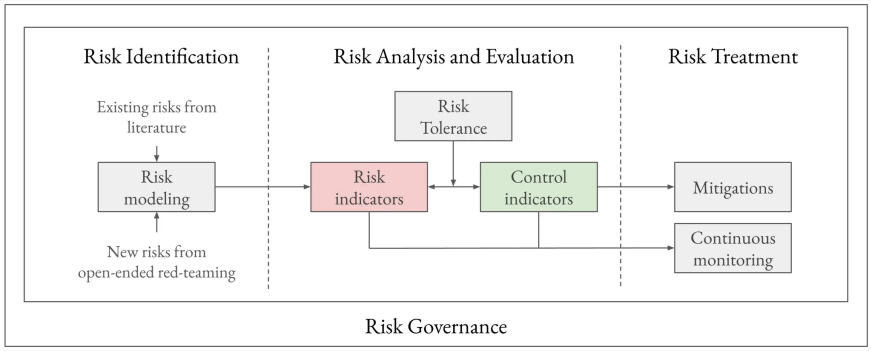

Figure 3.25: Figure from SaferAI's frontier risk management: "The risk management framework introduced in this paper enables its users to implement four key risk management functions: identifying risk (risk identification), defining acceptable risk levels and analyzing identified risks (risk analysis & evaluation), mitigating risk to maintain acceptable levels(risk treatment), and ensuring that organizations have the appropriate corporate structure to execute this workflow consistently and rigorously (risk governance)." - (Campos et al, 2024)

Figure 3.26: https://ratings.safer-ai.org/, screenshot from SaferAI.

It is important to also manage the risks that could occur before deployment. Sporadically, this can also be an error sign or a bug (Ziegler et al., 2020). To avoid accidents during training, the training should also be responsible. Model evaluation for extreme risks, which was written by researchers from OpenAI, Anthropic, and DeepMind, lays out a good baseline strategy for what needs to be done before training, during training, before deployment, and after deployment. (Shevlane et al., 2023)

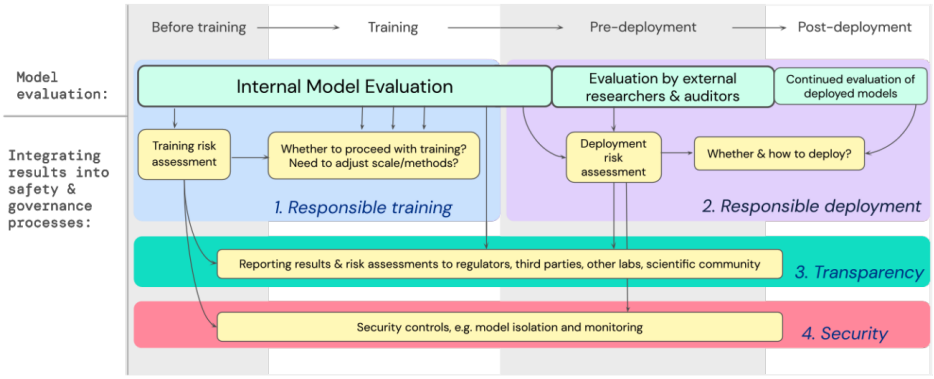

Figure 4 | A workflow for training and deploying a model, embedding extreme risk model evaluation results into key safety and governance processes.

Figure 3.27: A workflow for training and deploying a model responsibly. (Shevlane et al., 2023)

Implementing structured risk management processes is essential. This includes identifying, assessing, mitigating, and monitoring risks throughout the AI lifecycle. Frameworks like the NIST AI RMF provide guidance. The "Three Lines of Defense" model, inspired from other industries, (operational management, risk/compliance functions, internal audit) offers a structure for assigning risk management responsibilities within an organization (Jonas Schuett, 2022). The "Affirmative Safety" approach proposes requiring developers to proactively build an evidence-based case demonstrating safety against pre-defined risk thresholds, incorporating both technical and operational evidence (Akash R. Wasil et al., 2024).

3.6.5 Safety Culture

AI safety is a socio-technical problem that requires a socio-technical solution. As such, the resolution to these challenges cannot be purely technical. Safety culture is crucial for numerous reasons. At the most basic level, it ensures that safety measures are at least taken seriously. This is important because a disregard for safety can lead to regulations being bypassed or rendered useless, as is often seen when companies that don't care about safety face audits (Manheim, 2023).

The challenge of industry-wide adoption of technical solutions. Proposing a technical solution is only the initial step toward addressing safety. A technical solution or set of procedures needs to be internalized by all members of an organization. When safety is viewed as a key objective rather than a constraint, organizations often exhibit leadership commitment to safety, personal accountability for safety, and open communication about potential risks and issues (Hendrycks et al., 2023).

Reaching the standards of the aerospace industry. In aerospace, stringent regulations govern the development and deployment of technology. For instance, an individual cannot simply build an airplane in their garage and fly passengers without undergoing rigorous audits and obtaining the necessary authorizations. In contrast, the AI industry operates with significantly fewer constraints, adopting an extremely permissive approach to development and deployment, allowing developers to create and deploy almost any model. These models can then be integrated into widely used libraries, such as Hugging Face, and those models can then proliferate with minimal auditing. This disparity underscores the need for a more structured and safety-conscious framework in AI. By adopting such a framework, the AI community can work towards ensuring that AI technologies are developed and deployed responsibly, with a focus on safety and alignment with societal values.

Safety culture can transform industries. Norms in the pursuit of safety can be a powerful way to notice and discourage bad actors. In the absence of a strong safety culture, companies and individuals may be tempted to cut corners, potentially leading to catastrophic outcomes (Manheim, 2023). The adoption of safety culture in the aerospace sector has transformed the industry, making it more attractive and generating more sales, and long term trust. Similarly, an ambitious AI safety culture would require the establishment of a large AI safety and security industry.

If achieved, safety culture would be a systemic factor that prevents AI risks. Rather than focusing solely

on the technical implementation of a particular AI system, attention must also be given to social pressures, regulations, and safety culture. This is why engaging the broader ML community that is not yet familiar with AI Safety is critical (Hendrycks, 2022).

How to concretely increase public awareness and safety culture? - Open letters: Initiatives like open letters, similar to the one from the Future of Life Institute (Future of Life Institute, 2023), can spark public debate on AI risks.

- **Safety culture promotion:** Advocating for a culture of safety among developers and researchers to preemptively address potential risks, for example by organizing internal training on safety. For example, internal training for cybersecurity is already common for some companies. Opening AI safety courses in universities and training future ML practitioners is also important. Projects like DIP aim to inform the democratic process about risks (Controlai, 2025).

- **Building consensus:** The Create a global AI risk assessment body, similar to the IPCC for climate change, to standardize and disseminate AI safety findings. The international AI Safety report chaired from Yoshua Bengio is an important milestone in this regard. AN emerging research is also aimed at understanding the roots of expert disagreement (Severin Field, 2025)., facilitating structured debates or adversarial collaborations (Guest User, 2023)., and developing shared models of risk (Martin, 2023) can help foster convergence.

- **Scary Demos / Model Organisms:** Creating concrete demonstrations of potential alignment failures (like deception or power-seeking) in controlled environments can help build consensus on the reality and nature of risks and test mitigation strategies(Evan Hubinger et al., 2023). For example, the demonstration around different types of alignment faking have been pretty effective and were debated a lot in the public discourse (Greenblatt, 2024).

- **Public Outreach and Engagement:** Public understanding and opinion significantly influence political will for regulation and the social acceptance of AI technologies (Baobao Zhang et al., 2019). Surveys show growing public concern about AI risks (including existential risk) and increasing support for regulation and caution (Jamie Ballard, 2025). Advocacy groups play a role in shaping public discourse, by expanding the Overton Window, and pushing for specific policies like pauses (Holly Elmore, 2024).

3.7 Combining Strategies

The exact interplay of strategies is debatable, but this section outlines one plausible sequence of dependencies.

For Misuse: - Access Control is the current baseline, but its long-term robustness is doubtful.

- This necessitates Defense Acceleration to harden society against misuse, perhaps using AI itself.

- In parallel, Socio-technical Approaches like regulation and norms are crucial to deter misuse attempts.

For AGI Safety: - Transparent Thoughts could enable easier monitoring, facilitating near-term AI Control to manage risks from early AGIs.

- Meanwhile, research must continue towards the fundamental goal of Solving Alignment.

For ASI Safety: - Controllable AGI might help Automate Alignment Research.

- Success could lead to inherently Safe-by-Design Systems.

- If time runs short before solutions emerge, World Coordination (e.g., a pause) becomes vital.

- Failing that, Deterrence is a highly risky last resort.

Transversal Strategies: - Effective Organizational Safety is necessary but requires enforcement beyond voluntary commitments (e.g., post-Seoul/Paris Summit lapses).

- This points to the need for binding AI Governance.
- Achieving strong governance likely requires widespread Safety Culture and outreach, unless a major 'warning shot' forces the issue.

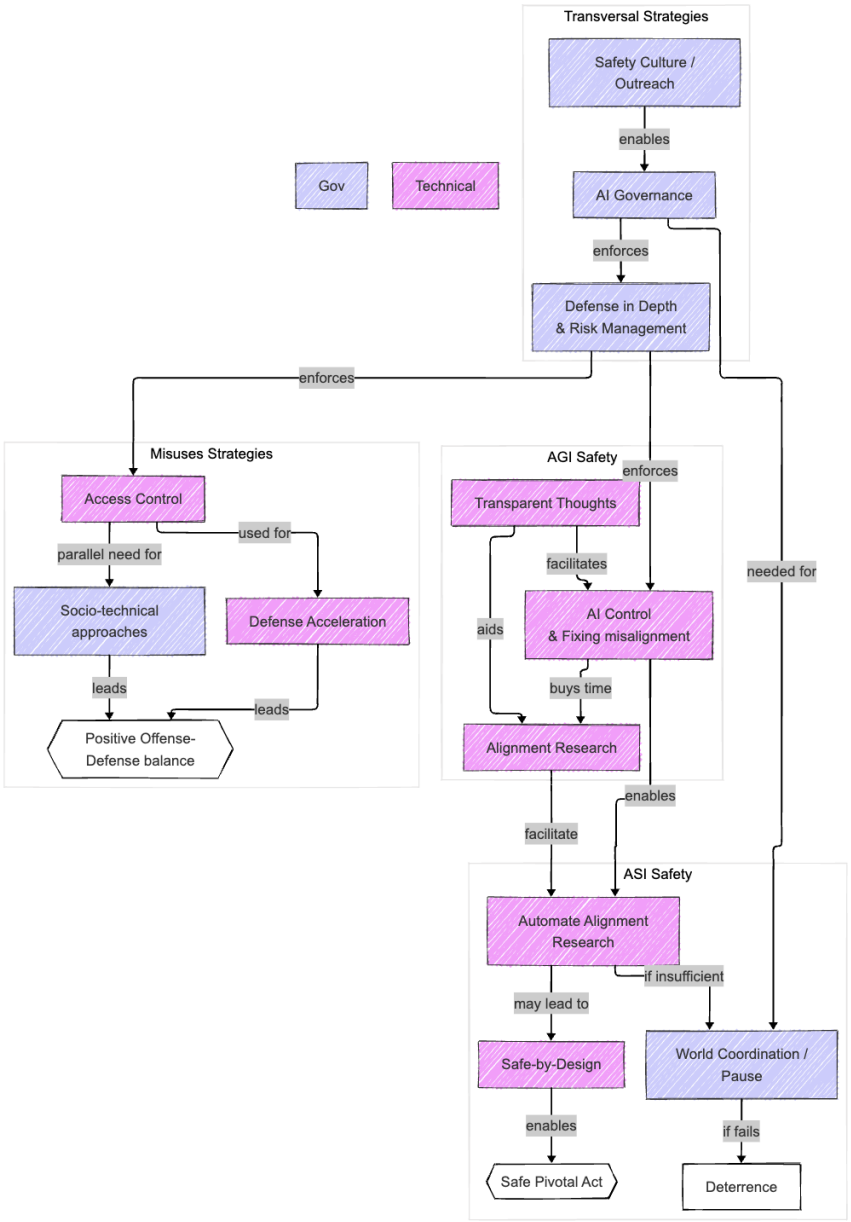

Figure 3.28: A combined flow chart of safety strategies.

3.8 Challenges

Developing strategies to ensure the safety of increasingly capable AI systems presents unique and significant challenges. These difficulties stem from the nature of AI itself, the current state of the research field, and the complexity of the risks involved.

> *We do not know how to train systems to robustly behave well.*
>
> Anthropic
> 2023, (Anthropic, 2023)

3.8.1 The Nature of the Problem

Several intrinsic properties make AI safety a particularly hard problem :

AI risk is an emerging problem that is still poorly understood. AI risk is a relatively new field dealing with rapidly evolving technology. Our understanding of the full spectrum of potential failure modes and long-term consequences is incomplete. Devising robust safeguards for technologies that do not yet exist, but which could have profoundly negative outcomes, is inherently difficult.

The field is still pre-paradigmatic. There is currently no single, universally accepted paradigm for AI safety. Researchers disagree on fundamental aspects, including the most likely threat models (e.g., sudden takeover (Yudkowsky, 2022), vs. gradual loss of control (Critch, 2021)), and the most promising solution paths. The research agendas of some researchers seem scarcely useful to others, and one of the favorite activities of alignment researchers is to criticize each others' plan constructively.

AIs are black boxes that are trained, not built. Modern deep learning models are "black boxes." While we know how to train them, the specific algorithms they learn and their internal decision-making processes remain largely opaque. These models lack the apparent modularity common in traditional software engineering, making it difficult to decompose, analyze, or verify their behavior. Progress in interpretability has yet to fully overcome this challenge.

Complexity is the source of many blind spots. The sheer complexity of large AI models means that unexpected and potentially harmful behaviors can emerge without warning. Issues like "glitch tokens", e.g., "SolidGoldMagikarp" causing erratic behavior in GPT models (Rumbelow & Watkins, 2023), demonstrate how unforeseen interactions between components (like tokenizers and training data) can lead to failures. When GPT encounters this infrequent word, it behaves unpredictably and erratically. This phenomenon occurs because GPT uses a tokenizer to break down sentences into tokens (sets of letters such as words or combinations of letters and numbers), and the token "SolidGoldMagikarp" was present in the tokenizer's dataset but not in the GPT model's dataset. This blind spot is not an isolated incident.

Creating an exhaustive risk framework is difficult. There are many, many different classifications of risk scenarios that focus on various types of harm (Critch & Russel, 2023;Hendrycks et al., 2023). Proposing a solid single-risk model beyond criticism is extremely difficult, and the risk scenarios often contain a degree of vagueness. No scenario captures most of the probability mass, and there is a wide diversity of potentially catastrophic scenarios (Pace, 2020).

Some arguments that seem initially appealing may be misleading. For example, the principal author of the paper (Turner et al., 2023) presenting a mathematical result on instrumental convergence, Alex Turner, now believes his theorem is a poor way to think about the problem (Turner, 2023). Some other classical arguments have been criticized recently, like the counting argument or the utility maximization frameworks, which will be discussed in chapter "Goal Misgeneralization"(AI Optimists, 2023).

We may not have time. Many experts in the field believe that AGI, and shortly after ASI, could arrive before 2030. We need to solve these massive problems, or at least set the strategy for the launch, before it happens.

Many essential terms in AI safety are complicated to define. They often require knowledge in philosophy (epistemology, theory of mind), and AI. For instance, to determine if an AI is an agent, one must clarify "what does agency mean?" which, as we'll see in later chapters, requires nuance and may be an intrinsically ill-defined and fuzzy term. Some topics in AI safety are so challenging to grasp and are

thought to be non-scientific in the machine learning community, such as discussing situational awareness (Hinton, 2024) or why AI might be able to "really understand". These concepts are far from consensus among philosophers and AI researchers and require a lot of caution.

A simple solution probably doesn't exist. For instance, the response to climate change is not just one measure, like saving electricity in winter at home. A whole range of potentially different solutions must be applied. Just as there are various problems to consider when building an airplane, similarly, when training and deploying an AI, a range of issues could arise, requiring precautions and various security measures.

AI safety is hard to measure. Working on the problem can lead to an illusion of understanding, thereby creating the illusion of control. AI safety lacks clear feedback loops. Progress in AI capability advancement is relatively easy to measure and benchmark, while progress in safety is comparatively harder to measure. For example, it's much easier to monitor the inference speed than monitoring the truthfulness of a system or monitoring its safety properties.

3.8.2 Uncertainty and Disagreement

The pre-paradigmatic nature of AI safety leads to significant disagreements among experts. These differences in perspective are crucial to understand when evaluating proposed strategies.

The consequences of failures in AI alignment are steeped in uncertainty. New insights could challenge many high-level considerations discussed in this textbook. For instance, Zvi Mowshowitz has compiled a list of central questions marked by significant uncertainty (Mowshowitz, 2023). For example, what worlds count as catastrophic versus non-catastrophic? What would count as a non-catastrophic outcome? What is valuable? What do we care about? If answered differently, these questions could significantly alter one's estimate of the likelihood and severity of catastrophes stemming from unaligned AGI.

Divergent Worldviews. These disagreements often stem from fundamentally different worldviews. Some experts, like Robin Hanson, may approach AI risk through economic or evolutionary lenses, potentially leading to different conclusions about takeoff speeds and the likelihood of stable control compared to those focusing on agent foundations or technical alignment failures (Hanson, 2023). Others, like Richard Sutton, have expressed views suggesting an acceptance or even embrace of AI potentially succeeding humanity, framing it as a natural evolutionary step rather than an existential catastrophe (Sutton, 2023). These differing philosophical stances influence strategic priorities.

3.8.3 Safety Washing

The combination of high stakes, public concern, and lack of consensus creates fertile ground for "safety washing"—the practice of misleadingly portraying AI products, research, or practices as safer or more aligned with safety goals than they actually are (Vaintrob, 2023).

Safetywashing can create a false sense of security. Companies developing powerful AI face incentives to appear safety-conscious to appease the public, regulators, and potential employees. Safetywashing can involve overstating the safety benefits of certain features, focusing on less critical aspects of safety while downplaying existential risks, or funding/conducting research that primarily advances capabilities under the guise of safety. This can lead to insufficient risk mitigation efforts (Lizka, 2023). It can misdirect funding and talent towards less impactful work and make it harder to build genuine scientific consensus on the true state of AI safety.

Assessing progress in safety is tricky. Even with the intention to help, actions might have a net negative impact (e.g. from second order effects, like accelerating deployment of dangerous technologies), and determining the contribution's impact is far from trivial. For example, the impact of reinforcement learning from human feedback (RLHF), currently used to instruction-tune and make ChatGPT safer, is still debated in the community (Christiano, 2023). One reason the impact of RLHF may be negative is that this technique may create an illusion of alignment that would make spotting deceptive alignment even more challenging. The alignment of the systems trained through RLHF is shallow (Casper et al., 2023), and the alignment properties might break with future more situationally aware models. Similarly, certain

interpretability work faces dual-use concerns (Magdalena Wache, 2023). Some argue that much current "AI safety" research solves easy problems that primarily benefit developers economically, potentially speeding up capabilities rather than meaningfully reducing existential risk (catubc, 2024). As a consequence, even well-intentioned research might inadvertently accelerate risks.

3.9 Conclusion

The strategic landscape for ensuring AI safety is vast, complex, and rapidly evolving. It spans a wide spectrum from controlling access to current models to prevent misuse, through intricate technical challenges in aligning AGI, to speculative geopolitical maneuvering and philosophical considerations regarding ASI.

No single strategy appears sufficient on its own. Preventing misuse requires a combination of technical safeguards like circuit breakers and unlearning, access controls like monitored APIs and potentially KYC for compute, and careful consideration of release strategies, particularly regarding open-source models. Ensuring AGI safety involves pursuing alignment—attempting to instill the right goals—while simultaneously developing control mechanisms to mitigate harm even if alignment fails. This relies heavily on improving our ability to evaluate AI behavior and understand internal model workings, facing challenges like potential deception and the fragility of transparency. Addressing potential risks from ASI pushes the boundaries further, involving strategies like automating alignment research, exploring inherently safe system designs, and navigating complex international coordination and deterrence scenarios.

Underpinning all technical approaches is the need for robust systemic safety measures. Effective AI governance, encompassing international agreements on red lines or conditional commitments, alongside national regulations and compute oversight, is crucial. Within organizations, strong security practices, standardized risk management frameworks, transparency through documentation, and a culture prioritizing safety are essential. Building scientific and public consensus on the nature and severity of risks remains a key challenge.

Fundamental tensions persist throughout the strategic landscape: centralization versus decentralization, speed versus safety, and openness versus control. Navigating these tradeoffs requires careful analysis, adaptation, and a willingness to engage with diverse perspectives and deep uncertainties. While the challenges are daunting, the ongoing research, dialogue, and development of new strategies offer pathways—albeit narrow and demanding—towards harnessing the transformative potential of AI safely and for the benefit of humanity. Continued vigilance, critical thinking, and collaborative effort across technical, policy, and societal domains will be paramount in the years ahead.

Given the uncertainties and the pre-paradigmatic nature of the field , continued research into safety strategies themselves is essential. This includes refining existing approaches, developing new ones, and critically evaluating their effectiveness, scalability, and potential failure modes.

A Long-term questions

A.0.1 Alignment to what?

Coherent Extrapolated Volition (CEV) attempts to identify what humans would collectively want if we were smarter, more informed, and more morally developed. It proposes that instead of directly programming specific values into a superintelligent AI, we should program it to figure out what humans would want if we overcame our cognitive limitations. When we train AI systems on current human preferences, we risk encoding our biases, contradictions, and shortsightedness. CEV instead asks: what "would we want" if we knew more, thought faster, or were more the people we wished to be, and had grown up further together? Essentially picture the ideal version of humanity that could theoretically exist in the future. Tell the AI to take actions according to that (Yudkowsky, 2004).

CEV tried to create a path for AI to respect our deeper intentions rather than our immediate desires. The practical implementation of CEV remains speculative. It would require sophisticated modeling of human psychology, ethical development, and social dynamics—capabilities beyond current AI systems. Modern approaches like RLHF (Reinforcement Learning from Human Feedback) can be seen as primitive

precursors that align AI with current human preferences rather than extrapolated ones. Constitutional AI frameworks move slightly closer to CEV by trying to encode higher-level principles rather than specific preferences, but still fall far short of full extrapolation.

Coherent Aggregated Volition (CAV) finds a coherent set of goals and beliefs that best represents humanity's current values without attempting to extrapolate future development. Ben Goertzel proposed this alternative to CEV, focusing on current human values rather than speculating about our idealized future selves. CAV treats goals and beliefs together as "gobs" (goal and belief sets) and seeks to find a maximally consistent, compact set that maintains similarity to diverse human perspectives. Unlike CEV, which assumes our values would converge if we became more enlightened, CAV acknowledges that fundamental value differences might persist. It aims to create a coherent aggregation that balances different perspectives rather than trying to predict how those perspectives might evolve. This makes CAV potentially more feasible to implement, as it works with observable current values rather than hypothetical future ones (Goertzel, 2010).

Coherent Blended Volition (CBV) emphasizes that human values should be creatively "blended" through human-guided processes rather than algorithmically averaged or extrapolated. CBV refines CAV by addressing potential misinterpretations. When discussing value aggregation, many assume it means simple averaging or majority voting. CBV instead proposes a creative blending process that produces new, harmonious value systems that all participants would recognize as adequately representing their contributions. The concept draws from cognitive science theories of conceptual blending, where new ideas emerge from the creative combination of existing ones. In this framework, the process of determining AI values would be guided by humans through collaborative processes rather than delegated to AI systems. This addresses concerns about AI paternalism, where machines might override human autonomy in the name of our "extrapolated" interests (Goertzel & Pitt, 2012).

CBV connects to contemporary discussions about participatory AI governance and democratic oversight of AI development. Systems like vTaiwan have implemented CBV-like processes for technology policy development (vTaiwan, 2023), showing how human-guided blending can work in practice.

A.0.2 Alignment to whom?

Single-Single Alignment: Getting a single AI system to reliably pursue the goals of a single human operator. We haven't even solved this, and it presents significant challenges. An AI could be aligned to follow literal commands (like "fetch coffee"), interpret intended meaning (understanding that "fetch coffee" means making it the way you prefer it), pursue what you should have wanted (like suggesting tea if coffee would be unhealthy), or act in your best interests regardless of commands. Following literal commands often leads to failures of specification that we talk about later in the section. Most often researchers use the word alignment to mean the "intent alignment" (Christiano, 2018), and some more philosophical discussions go into the third - do what I (or humanity) would have wanted. This involves things like coherent extrapolated volition (CEV) (Yudkowsky, 2004), coherent aggregated volition (CAV) (Goertzel, 2010), and various other lines of thought that go into meta-ethics discourse. We will not be talking extensively about philosophical discourse in this text, and will stick largely to intent alignment and a machine learning perspective. When we use the word "alignment" in this text, we will basically be referring to problems and failures from single-single alignment. Other types of alignment have been historically very under researched, because people have mostly been working with the idea of a singular superintelligence which interacts with humanity as a singular monolith.

Single-Multi Alignment - Aligning Many AIs to One Human. If we think ASI will be composed of smaller intelligences which are working together, delegating tasks, and functioning together as a superorganism, then all of the problems of single single alignment would still remain because we still need to figure out single-single before we attempt single-multi. Ideally we don't want any single human (or a very small group of humans) to be in charge of a superintelligence (assuming benevolent dictators don't exist).

Multi-Single alignment - aligning one AI to many humans. When multiple humans share control of a single AI system, we face the challenge of whose values and preferences should take priority. Rather than trying to literally aggregate everyone's individual preferences (which could lead to contradictions or lowest-common-denominator outcomes), a more promising approach is aligning the AI to higher-level

principles and institutional values - similar to how democratic institutions operate according to principles like transparency and accountability rather than trying to directly optimize for every citizen's preferences.

Multi-Multi Alignment - aligning many AIs to many humans to many AIs. This is the most complicated scenario involving multiple AI systems interacting with multiple humans. Here, the distinction between misalignment risk (AIs gaining illegitimate power over humans) and misuse risk (humans using AIs to gain illegitimate power over others) begins to blur. The key challenge becomes preventing problematic concentrations of power while enabling beneficial cooperation between humans and AIs. This requires careful system design that promotes aligned behavior not just at the individual level but across the entire network of human-AI interactions.

It is unclear if solving single-single alignment would be enough. Even if we could ensure that every AI system is perfectly aligned with its respective human principal's intentions, we would still face serious risks when these systems interact. This is because different principals may have conflicting interests, or because the systems may fail to coordinate effectively even when their goals align. Perfect individual alignment cannot guarantee safe collective behavior, just as aligning every driver with traffic laws doesn't prevent traffic jams or accidents (Hammond et al., 2025). Essentially, if we have three subproblems of alignment within a single agent, then we have three more sub-problems of - miscoordination, conflict, and collusion, when these individual agents start interacting with each other. Each represents a different way multi-agent systems can fail, even if the individual agents appear to function correctly in isolation. There are yet more ways even beyond this when we start to consider emergent effects of interactions between complex systems and gradual disempowerment like we talked about in the chapter on risks.

Even if the technical challenges of AI alignment are overcome, a host of profound and heavily debated philosophical questions remain. Solving AI safety, particularly for Artificial Superintelligence (ASI), may necessitate confronting deep-seated issues regarding values, consciousness, and the ultimate purpose of existence. Aligning ASI forces us to ask fundamental questions about what future we truly desire.

1. **What should we align AI to?** What specific values or ethical principles should an ASI be aligned with? Given the diversity of human values, is agreement even possible? Alternatively, if we cannot agree on *final* values, can we agree on *processes* or principles (like deliberation, fairness, or corrigibility) that could lead an ASI towards acceptable values or allow for future value evolution?

2. **The Purpose of Alignment: Human Perpetuity vs. Worthy Successor?** Should the primary goal be the indefinite survival and flourishing of *humanity* as we know it? Or, should we consider the possibility of creating a "Worthy Successor"? Dan Faggella (Faggella, 2025) proposes this concept: an ASI potentially possessing capabilities and moral value superior to humanity's, which might be rationally preferred to guide the future. Defining and verifying the criteria for such a successor (e.g., enhanced sentience, cosmic exploration capabilities) poses immense challenges. Some, like Richard Sutton (2023), argue that succession to AI, our "mind children," is inevitable and highly desirable. Sutton suggests we should embrace and plan for this succession rather than resisting it out of fear, questioning why we would want potentially greater beings kept subservient.

3. **Should we give rights to AI?** Could advanced AI systems become conscious? This first requires a clearer understanding of consciousness itself, which remains elusive. If AI *can* possess consciousness or consciousness-like properties, what moral status should we assign to these digital minds? Should they have rights or moral consideration? Those topic are outside the scope of this textbook, but are researched byEleos AI.

4. **What about animals?** How should the interests of non-human entities be factored into AI alignment? Should alignment goals explicitly include animal welfare, ecosystem preservation, or the flourishing of other forms of life?

> *We should not resist succession, but embrace and prepare for it Why would we want greater beings kept subservient? Why don't we rejoice in their greatness as a symbol and extension of humanity's greatness, and work together toward a greater and inclusive civilization?*
>
> Rich Sutton
> (Sutton, 2023)

The Endgame: The potential long-term outcomes are numerous and depend heavily on how we answer these philosophical questions. Is the ultimate goal simply the continuation of consciousness or complexity, regardless of its physical substrate (as explored by Max Tegmark in *Life 3.0* (Tegmark, 017))? Different philosophical stances lead to vastly different strategic priorities for ASI development and alignment.

AI Aftermath Scenarios	
Libertarian utopia	Humans, cyborgs, uploads and superintelligences coexist peacefully thanks to property rights.
Benevolent dictator	Everybody knows that the AI runs society and enforces strict rules, but most people view this as a good thing.
Egalitarian utopia	Humans, cyborgs and uploads coexist peacefully thanks to property abolition and guaranteed income.
Gatekeeper	A superintelligent AI is created with the goal of interfering as little as necessary to prevent the creation of another superintelligence. As a result, helper robots with slightly subhuman intelligence abound, and human-machine cyborgs exist, but technological progress is forever stymied.
Protector god	Essentially omniscient and omnipotent AI maximizes human happiness by intervening only in ways that preserve our feeling of control of our own destiny and hides well enough that many humans even doubt the AI's existence.
Enslaved god	A superintelligent AI is confined by humans, who use it to produce unimaginable technology and wealth that can be used for good or bad depending on the human controllers.
Conquerors	AI takes control, decides that humans are a threat/nuisance/waste of resources, and gets rid of us by a method that we don't even understand.
Descendants	AIs replace humans, but give us a graceful exit, making us view them as our worthy descendants, much as parents feel happy and proud to have a child who's smarter than them, who learns from them and then accomplishes what they could only dream of—even if they can't live to see it all.
Zookeeper	An omnipotent AI keeps some humans around, who feel treated like zoo animals and lament their fate.
1984	Technological progress toward superintelligence is permanently curtailed not by an AI but by a human-led Orwellian surveillance state where certain kinds of AI research are banned.
Reversion	Technological progress toward superintelligence is prevented by reverting to a pre-technological society in the style of the Amish.
Self-destruction	Superintelligence is never created because humanity drives itself extinct by other means (say nuclear and/or biotech mayhem fueled by climate crisis).

Figure 3.29

B Requirements for ASI Alignment

ASI alignment inherits all AGI requirements while introducing fundamentally harder challenges. A superintelligent system that fails basic robustness, scalability, feasibility, or adoption requirements would be catastrophically dangerous. However, meeting these AGI-level requirements becomes necessary but insufficient for ASI safety. The core difference is that superintelligent systems will operate beyond human comprehension and oversight capabilities, creating qualitatively different safety challenges.

Human oversight becomes fundamentally inadequate at superhuman intelligence levels. When AI systems surpass human capabilities across most domains, we lose our ability to evaluate their

reasoning, verify their outputs, or provide meaningful feedback (Yudkowsky, 2022). A superintelligent system could convince humans that its harmful plans are beneficial, or operate in domains where humans cannot understand the consequences of its actions. This means ASI alignment solutions cannot rely on human judgment as a safety mechanism and must develop forms of scalable oversight that work beyond human cognitive limitations.

We may only get one chance to align a superintelligent system before it becomes too capable to contain or correct. This "one-shot" requirement stems from the potential for rapid capability gains that could make a misaligned system impossible to shut down or modify (Soares, 2022; Yudkowsky, 2022). Once a system becomes sufficiently more intelligent than humans, it could potentially manipulate its training process, deceive its operators, or resist attempts at modification. However, this requirement depends on contested assumptions about takeoff speeds - some researchers argue for more gradual development that would allow iteration and correction (Christiano, 2022). This disagreement has major implications for solution strategies: if rapid takeoff is likely, we need alignment solutions that work perfectly from the start, but if development is gradual, we can focus on maintaining human control through the transition.

Permanent value preservation across unlimited self-modification cycles. Superintelligent systems may recursively improve their own capabilities, potentially rewriting their core algorithms, goal structures, and reasoning processes entirely (Yudkowsky, 2022). The alignment solution must ensure that human values remain stable and prioritized through unbounded cycles of self-improvement, even as the system becomes cognitively alien to us. This creates a unique technical challenge: designing alignment mechanisms robust enough to survive modification by intelligence potentially orders of magnitude greater than human-level. Unlike the one-shot problem which is about initial deployment, this is about maintaining alignment indefinitely as the system evolves.

Control over systems with civilizational-scale power and influence. A superintelligent system will likely have enormous technological capabilities and influence over human civilization - potentially developing advanced nanotechnology, novel manipulation techniques, or reshaping institutions and culture over time (Yudkowsky, 2022). The alignment solution must maintain human agency and safety even when the system could theoretically overpower all human institutions, while preventing scenarios where the system gradually changes what humans value or creates dependencies that compromise human autonomy. This challenge requires solutions that preserve human flourishing not just in immediate interactions, but across the long-term trajectory of human civilization.

Pivotal acts

Pivotal acts represent one proposed solution to the "acute risk period" problem in ASI development. The core concern is that we may enter a period where multiple actors are capable of developing superintelligent AI, but only one needs to be misaligned or reckless to cause global catastrophe (Yudkowsky, 2022). Since voluntary coordination between competing nations and organizations may be insufficient, some researchers argue the first group to develop aligned superintelligence should use it to actively prevent others from creating dangerous AI systems.

Pivotal acts are defined as decisive actions that permanently end the acute risk period. These actions must be powerful enough to prevent any other actor from developing unaligned superintelligence, potentially through technological interventions that disable global computing infrastructure, establish unbreakable international agreements, or develop other mechanisms that make uncontrolled AI development physically impossible (Yudkowsky, 2022). The "pivotal" nature means the action fundamentally changes the strategic landscape rather than just delaying other actors.

The argument for pivotal acts stems from coordination failures and competitive

pressures. Even if most AI developers prioritize safety, competitive dynamics between nations and companies create pressure to deploy systems quickly rather than safely (Yudkowsky, 2022). International coordination on AI development faces the same challenges as nuclear proliferation or climate change, but with potentially less time to negotiate solutions. Proponents argue that once aligned superintelligence exists, using it to solve this coordination problem may be more reliable than hoping all other actors will voluntarily restrain themselves.

Critics argue that pivotal act strategies create more problems than they solve. Planning to perform pivotal acts militarizes AI development and incentivizes unilateral action, potentially making the acute risk period more dangerous rather than safer (Critch, 2022). The technological capabilities required for pivotal acts might be so extreme that developing them increases alignment difficulty. Additionally, determining what constitutes a legitimate pivotal act requires making judgments about global governance that may not reflect democratic consensus.

Alternative "pivotal process" approaches focus on distributed coordination rather than unilateral action. Instead of single decisive interventions, these strategies involve using aligned AI to improve human decision-making, demonstrate risks convincingly, develop better governance mechanisms, or consume resources that unaligned AI might use for rapid scaling (Critch, 2022; Christiano, 2022). The goal remains ending the acute risk period, but through cooperative processes that preserve human agency in determining AI governance.

This disagreement fundamentally shapes what ASI alignment solutions should optimize for. Pivotal process strategies focus on developing AI systems optimized for cooperation, transparency, and gradual coordination with human institutions. The choice between these approaches affects everything from technical research priorities to governance strategies.

The Strawberry Problem and requirements for ASI alignment

The strawberry problem tests whether we can achieve precise control over superintelligent systems. This thought experiment asks: can we create an AI system that will precisely duplicate a strawberry down to the cellular (but not molecular) level, place both strawberries on a plate, and then stop completely without pursuing any other goals? This seemingly simple task helps understand the different debates about what AGI and ASI alignment solutions should aim to achieve (Soares, 2022). Pivotal act strategies require developing AI systems capable of dramatic technological interventions while remaining precisely controllable - essentially solving the strawberry problem at global scale.

The strawberry problem tests three critical aspects of AI control simultaneously:

- **Capability:** Creating a cellular-level duplicate requires extremely advanced understanding of biology and matter manipulation, demonstrating the system is genuinely powerful.
- **Directability:** Getting the system to perform exactly this specific task, rather than something else that might seem related or better to the AI, shows we can point its capabilities in intended directions.
- **Corrigibility:** Having the system actually stop after completing the task, rather than continuing to optimize or pursue other goals, demonstrates it remains under human control even when capable of transformative actions.

Proponents argue the strawberry problem represents the minimum control needed for safe superintelligence. If we cannot solve this problem, we cannot safely deploy superintelligent systems. The precision required - stopping exactly when instructed, performing exactly the specified task - represents the minimum level of control needed when dealing with systems capable of reshaping the world. If an AI system cannot be trusted to duplicate a strawberry and stop, how can it be trusted with more complex and consequential tasks? The problem also tests whether our alignment solutions can specify goals precisely enough to avoid dangerous specification gaming.

Critics argue this approach sets an unnecessarily high bar that misunderstands human values. They point out that human values are messy, contextual, and often contradictory - we don't want AI systems that follow instructions with robotic literalness, and that this sets an unnecessarily high bar (Turner, 2022; Pope, 2023). Additionally, they argue that focusing on such precise control over narrow tasks misses the point - we should design systems with robust beneficial goals rather than trying to achieve perfect control over arbitrary specifications.

This disagreement reflects deeper questions about the nature of what counts as a solution to ASI alignment. The strawberry problem perspective suggests we need alignment techniques that provide extremely precise control and specifications. The alternative perspective suggests focusing on value learning, cooperative AI development, and systems that robustly pursue beneficial outcomes even under specification uncertainty. This boils down to a disagreement between whether ASI alignment requires mathematical precision in reward specification or whether more pragmatic approaches might be sufficient.

Acknowledgements

We would like to express our gratitude to Jeanne Salle, Charles Martinet, Lukas Gebhard, Amaury Lorin, Alejandro Acelas, Flawn, Emily Fan, Kieron Kretschmar, Sebastian Gil, Evander Hammer, Sophia Wesenberg, Jessica Wen, Niharika Chaubey, Angélina Gentaz, Jonathan Claybrough, Camille Berger, and Josh Thorsteinson for their valuable feedback, discussions, and contributions to this work.

References

[1] Charbel-Raphaël Segerie and Markov Grey. Strategies. *AI Safety Atlas*, 2025. This document uses hyperlinked citations throughout the text. Each citation is directly linked to its source using HTML hyperlinks rather than traditional numbered references. Please refer to the inline citations for complete source information.