Chapter 3 - Solutions Landscape

3.1 Introduction

Although the field of AI safety is still in its infancy, several measures have already been identified that can significantly improve the safety of AI systems. While it remains to be seen if these measures are sufficient to fully address the risks posed by AI, they represent essential considerations. The diagram below provides a high-level overview of the main approaches to ensuring the safe development of AI.

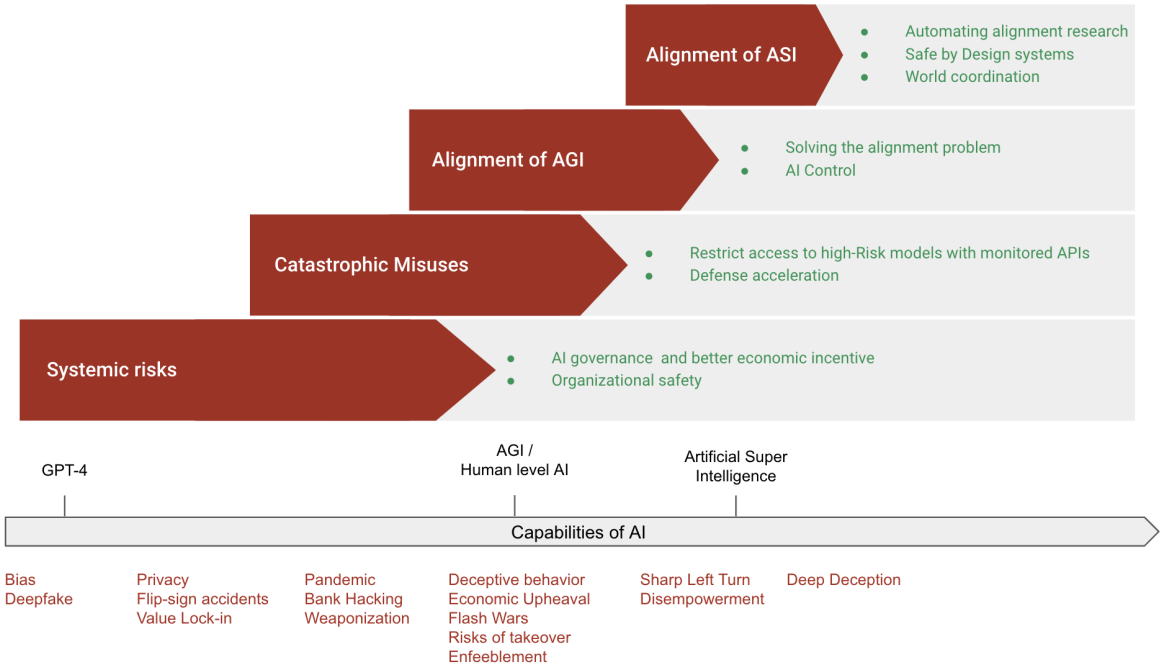

Figure: *Tentative diagram summarizing the main high-level approaches to make AI development safe.*

The information in this chapter is far from exhaustive and only scratches the surface of the complex landscape of AI safety. Readers are encouraged to explore this recent list of agendas for a more comprehensive review.

3.2 AI Safety is Challenging

Specific properties of the AI safety problem make it particularly difficult.

AI risk is an emerging problem that is still poorly understood. We are not yet familiar with all its different aspects, and the technology is constantly evolving. It's hard to devise solutions for a technology that does not yet exist, but setting up these guardrails are also necessary because the outcome can be very negative.

The field is still pre-paradigmatic. AI safety researchers disagree on the core problems, difficulties, and main threat models. For example, some researchers think that takeover risks are more likely (source), whereas other research emphasizes more progressive failure modes with progressive loss of control (source). Because of this, alignment research is currently a mix of different agendas that need more unity. The alignment agendas of some researchers seem incommensurate hopeless to others, and one of the favorite activities of alignment researchers is to criticize each other constructively.

AIs are black boxes that are trained, not built. We know how to train them, but we do not know which algorithm is learned by them. Without progress in interpretability, they are giant inscrutable matrices of numbers with little modularity. In software engineering, modularity helps break down software into simpler parts, allowing for better problem-solving. In deep learning models, modularity is almost nonexistent: to date, interpretability has failed to decompose a deep neural network into modular structures (source). As a result, behaviors exhibited by deep neural networks are not understood and keep surprising us.

Complexity is the source of many blind spots. New failure modes are frequently discovered. For example, issues arise with glitch tokens, such as "SolidGoldMagikarp" (source). 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. For example, on the day Microsoft's Tay chatbot, BingChat, or ChatGPT were launched, the chatbots were poorly tuned and exhibited many new emerging undesirable chatbot behaviors. Finding solutions when you don't know there is a problem is hard.

Creating an exhaustive risk framework is difficult. There are many, many different classifications of risk scenarios that focus on various types of harm (source)(source). 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 (source).

Some arguments or frameworks that seem initially appealing may be flawed and misleading. For example, the principal author of the paper (source) presenting a

mathematical result on instrumental convergence, Alex Turner, now believes his theorem is a poor way to think about the problem (source). Some other classical arguments have been criticized recently, like the counting argument or the utility maximization frameworks, which will be discussed in the following chapters. (source)

Intrinsic fuzziness. Many essential terms in AI safety are complicated to define, requiring 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 (ML) community, such as discussing *situational awareness* (source) 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 very 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.

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 (source). 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 (source), and the alignment properties might break with future more situationally aware models. Another worry is that many speculative failure modes appear only with more advanced AIs, and the fact that systems like GPT-4 can be instruction-tuned might not be an update for the risk models that are the most worrying.

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 easy to measure and benchmark, while progress in safety is comparatively much 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.

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 critical questions marked by significant uncertainty and strong disagreements both ethical and technical (source). 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? (source) If answered differently, these questions could significantly alter one's estimate of the likelihood and severity of catastrophes stemming from unaligned AGI. Diverse responses to these critical questions highlight why individuals familiar with the alignment risks often hold differing opinions. For example, figures like Robin Hanson and Richard Sutton suggest that the concept of losing control to AIs might not be as dire as it seems. They argue there is little difference between nurturing a child who eventually influences the economy and developing AI based on human behavior that subsequently assumes economic control (source)(source).

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

3.3 Definitions

AI alignment is hard to define but generally refers to the process and goal of ensuring that an AI system's behaviors and decision-making processes are in harmony with the intentions, values, and goals of its human operators or designers. The main concern is whether the AI is pursuing the objectives we want it to pursue. AI alignment does not encompass systemic risks and misuses.

AI safety, on the other hand, encompasses a broader set of concerns about how AI systems might pose risks or be dangerous and what can be done to prevent or mitigate those risks. Safety is concerned with ensuring that AI systems do not inadvertently or deliberately cause harm or danger to humans or the environment.

AI ethics is the field that studies and formulates the moral principles guiding the development and use of AI systems to ensure they respect fundamental human values. It addresses issues such as bias, transparency, accountability, privacy, and societal impact, with the goal of developing trustworthy, fair, and beneficial AI through multidisciplinary collaboration and appropriate governance frameworks. This chapter is mostly not focusing on AI ethics.

AI control is about the mechanisms that can be put in place to manage and restrain an AI system if it acts contrary to human interests. Control might require physical or virtual

constraints, monitoring systems, or even a kill switch to shut down the AI if necessary (source).

In essence, alignment seeks to prevent preference divergence by design, while control deals with the consequences of potential divergence after the fact (source). The distinction is important because some approaches to the AI safety problem focus on building aligned AI that inherently does the right thing, while others focus on finding ways to control AI that might have other goals. For example, a research agenda from the research organization Redwood Research argues that even if alignment is necessary for superintelligence-level AIs, control through some form of monitoring may be a working strategy for AIs in the human regime (source).

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.

In the remainder of this chapter, we will reuse the classification of risks we used in the last chapter: misuse, alignment, and systemic risks.

3.4 Misuse

Recap on the potential misuses of AI, from present threats to future ones:

- **Misinformation campaigns:** Spreading false narratives or manipulating elections.
- **Deep fakes:** Misleading people using fake videos or calls that seem real
- **Loss of privacy:** Intercepting personal data or eavesdropping on people's activities.
- **Destructive AI conflicts:** Using AI to create more lethal and autonomous weaponry.
- **Cyberattacks:** Targeting critical structures such as major financial institutions or nuclear facilities using an AI's superhuman abilities.
- **Engineered pandemics:** Designing and producing deadly pathogens.

In addressing the mitigation of ***future*** potential misuses of mighty AI, two distinct strategies emerge:

- **Strategy A: Limiting the proliferation of high-risk models - for example, via monitored APIs.** The Monitored APIs strategy focuses on controlling access to AI models that could pose extreme risks, such as those capable of enhancing **cyberattacks** or facilitating the creation of **engineered pandemics**. By placing high-risk models behind monitored APIs, we ensure only authorized users can access the technology.

This approach is akin to digital containment, where the potential of AI to be weaponized or misused is curtailed by stringent access controls. This method also allows for detailed tracking of how the AI models are being used, enabling the early detection of misuse patterns.

- **Strategy B: "vaccinating" the society and "preparing vaccine factories" for many types of harm - a strategy called Defense Acceleration.** The Defense Acceleration (d/acc) strategy (source) advocates for the rapid development and deployment of defensive technologies. This strategy is premised on the belief that bolstering our defensive capabilities can outpace and mitigate the threats posed by offensive AI uses, such as cyberattacks or engineered biological threats. The essence of d/acc is to create a technological environment where the cost and complexity of offensive actions are significantly increased, thereby reducing their feasibility and attractiveness.

To address the potential misuse of ***current*** types of systems:

- **Strategy C: When problematic models are already widely available, one of the few solutions is to make it illegal to use them for clearly bad purposes.** This may include things like non-consensual deepfake sexual content, since there is no easy technical defense against these threats.

3.4.1 STRATEGY A: MONITORED APIS

As AGI becomes more accessible, easier to build and potentially destructive, we need as much control and monitoring as possible over who can use dangerous AIs. A preventative measure against misuse involves **restricting access to powerful models** capable of causing harm. This means placing high-risk models behind application programming interfaces (APIs) and monitoring their activity.

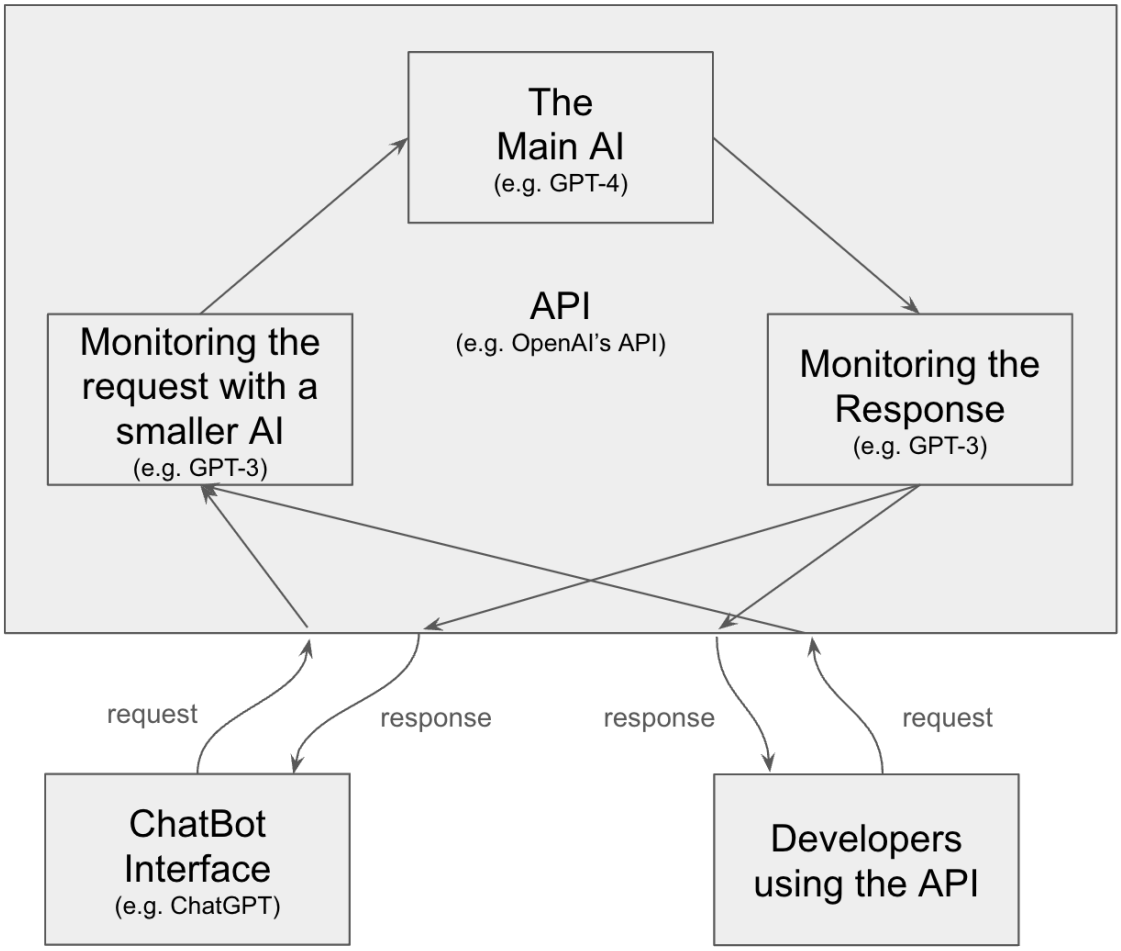

Figure: *This schema illustrates how an API works. This diagram is very simplified and is for illustration purposes only. OpenAI's API does not work like this.*

The necessity of model evaluation. The first step in this strategy is to identify which models are considered dangerous and which are not via model evaluation. The paper Model Evaluation for Extreme Risks [s], which was partly presented during the last chapter, lists a few critical, dangerous capabilities that need to be assessed.

Models classified as potentially dangerous should be monitored. The most hazardous AIs should be subject to careful controls. AIs with capabilities in biological research, cyber threats, or autonomous replication and adaptation should have strict access controls to prevent misuse for terrorism, similar to firearm regulation. These capabilities should be excluded from AIs designed for general purposes, possibly through dataset filtering or, more speculatively, through model unlearning (source). Dangerous AI systems providers should only allow controlled interactions through cloud services (source). It's also crucial to consider alternatives to the binary approach of entirely open or closed model sharing.

Strategies such as gradual and selective staged sharing, which allows time between model releases to conduct risk and benefit analyses as model sizes increase, could balance benefits against risks (source). Monitoring sensitive models behind APIs with anomaly detection algorithms could also be helpful.

A key monitoring strategy is the Know Your Customer (KYC) standard. This is a mandatory process adopted by companies and banks that involves verifying the identities of their clients or legal entities in line with current customer due diligence regulations. KYC concretely means requiring an identity card plus a background check before any services can be used. This is important for tracing malicious users.

Red Teaming can help assess if these measures 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.

The measures outlined above are the most straightforward to implement. A more detailed description of simple measures for preventing misuse is available here, and they appear to be technically feasible. It requires the willingness to take precautions and to place models behind APIs.

Dangerous models should not be hacked and exfiltrated. Research labs developing cutting-edge models must implement rigorous cybersecurity measures to protect AI systems from theft by outside actors and use sufficient cybersecurity defenses to limit proliferation. This seems simple, but it's not, and protecting models from nation-state actors could require extraordinary effort (source).

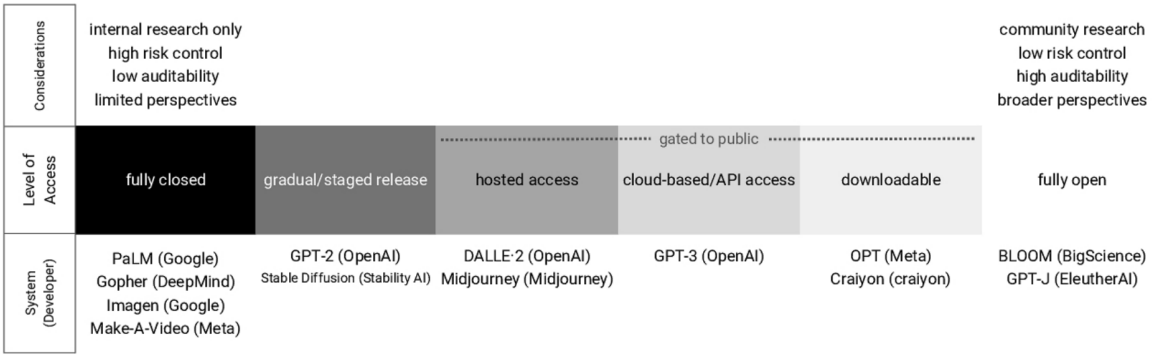

Figure: *Gradient of system access (source)*. *

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. Moreover, recent 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. (source) This is why placing models behind APIs makes sense, as it prevents unauthorized fine-tuning without the company's consent.

While promising to limit extreme risks like cyberattacks or biorisks, monitored APIs may not be as effective against the subtler threats of deep fakes and privacy erosion. Deep fakes, for instance, require less sophisticated AI models that are already widely available, and those models might not be classified as high-risk and, hence, not placed behind monitored APIs. More on this in strategy C.

> **Anthropic (source))**
>
> *[...] safeguards such as Reinforcement Learning from Human Feedback (RLHF) or constitutional training can almost certainly be fine-tuned away within the specified 1% of training cost.*

3.4.2 STRATEGY B: DEFENSE ACCELERATION

The above framework assumes that dangerous AIs are closed behind APIs and require a certain amount of centralization.

However, centralization can also pose systemic risks (source)**.** There's a trade-off between securing models behind APIs to control misuse and the risks of over-centralization (source). For instance, if in 20 years all companies worldwide rely on a single company's API, significant risks of value lock-in or fragility could arise because the whole world would be dependent on the political opinion of this model and its stability or instability of this API could be a single point of failure, without talking about power concentrations.

Another possible paradigm is that AIs should be open and decentralized. Yes, if models are open-sourced, we have to acknowledge that not all AIs will be used for good, just as we have to acknowledge that there are disturbed individuals who commit horrific acts of violence. Even if AIs are instruction-tuned before open-sourcing, it's possible to remove security barriers very easily (source), as we've seen earlier. This means that some people will misuse AI, and we need to prepare for that. For example, we would need to create more defenses in existing infrastructures. An example of defense would be to use models to iterate on all the world's open-source code to find security flaws so that good actors rather than malicious actors find security flaws. Another example would be to use the model to find holes in the security of the bioweapons supply chain and correct those problems.

Defense acceleration. Defense acceleration, or d/acc, is a framework popularized by Vitalik Buterin (source). d/acc is a strategic approach focusing on promoting technologies and innovations that inherently favor defense over offense. This strategy emphasizes developing and accelerating technologies that protect individuals, societies, and ecosystems from various threats rather than technologies that could be used for aggressive or harmful purposes. It's like vaccinating society against the potential downsides of our advancements. d/acc would also be a plausible strategy for maintaining freedom and privacy. It's a bit like ensuring everyone has a strong enough lock on their doors; if breaking in is tough, there's less chaos and crime. This is crucial for ensuring that technological advancements don't lead to dystopian scenarios where surveillance and control are rampant.

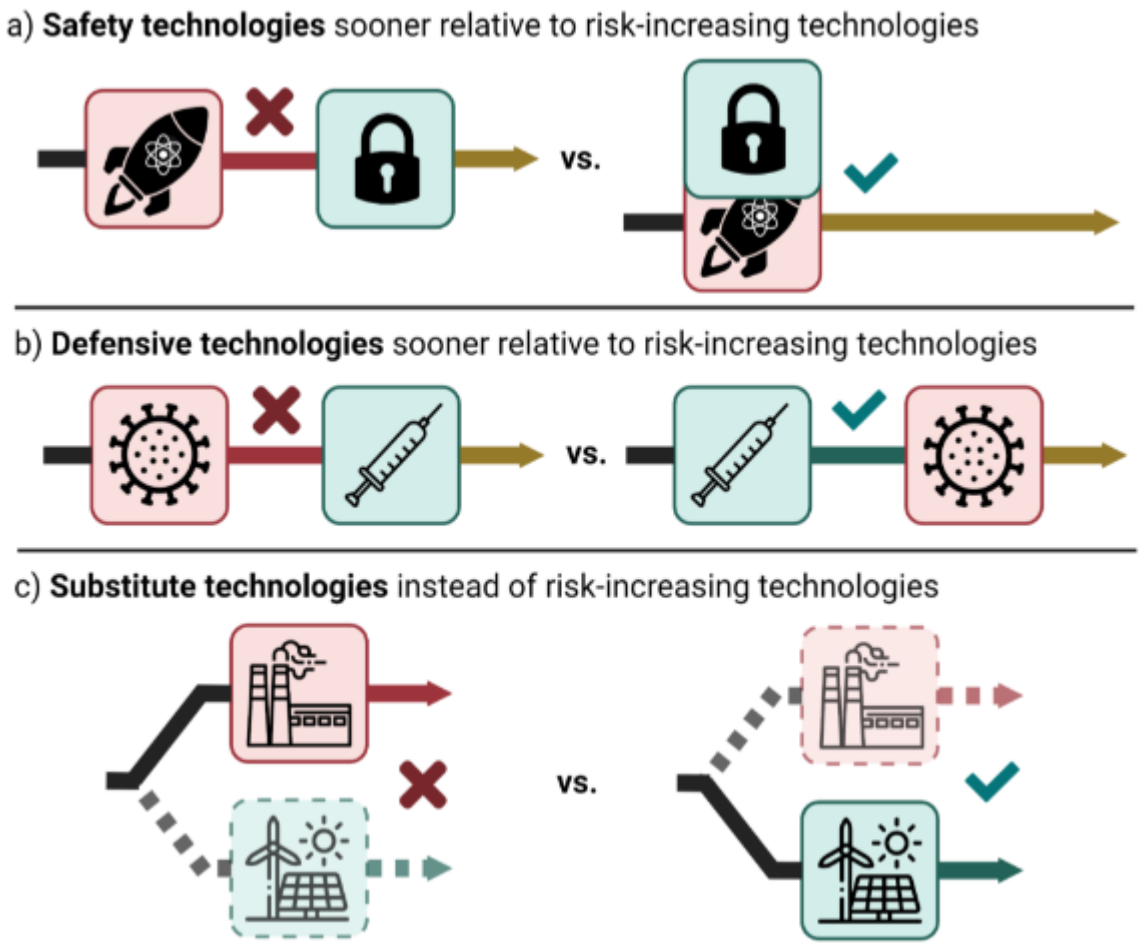

Figure: Mechanisms by which differential technology development can reduce negative societal impacts. (source)

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

A key consideration for the feasibility of the d/acc framework is the offense-defense balance: how hard is it to defend against an attacker? This concept is crucial to assess which models are more likely to be beneficial than detrimental if we open-source them. In traditional software development, open sourcing often shifts the offense-defense balance positively: the increased transparency allows a broader community of developers to identify and patch vulnerabilities, enhancing overall security (source). However, this dynamic could change with the open-sourcing of frontier AI models because they introduce new emerging types of risks that could not simply be patched. This represents a significant shift from traditional software vulnerabilities to more complex and dangerous threats that cannot be easily countered with defensive measures.

In the case of AI, sharing the most potent models may pose extreme risks that could outweigh the usual benefits of open-sourcing. For example, just as sharing the procedure for making a deadly virus seems extremely irresponsible, so too should freely distributing AI models that provide access to such knowledge.

The current balance for sharing frontier models remains uncertain; it has been clearly positive so far, but deploying increasingly powerful models could tip this balance towards unacceptable levels of risk.[1]

The dangers emerging from frontier AI are nascent, and the harms they pose are not yet extreme. That said, as we stand at the dawn of a new technological era with increasingly capable frontier AI, we are seeing signals of new dangerous capabilities. We must pay attention to these signals. Once more extreme harms start occurring, it might be too late to start thinking about standards and regulations to ensure safe model release. It is essential to exercise caution and discernment now.

The d/acc philosophy requires more research, as it's not clear that the offense-defense balance is positive before open-sourcing dangerous AIs, as open-sourcing is irreversible.

For a short review of different positions on open source, we recommend reading [2308.14253] The Promise and Peril of Artificial Intelligence.

Ajeya Cotra

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.

3.4.3 STRATEGY C: ADDRESSING RISKS FROM CURRENT AIS

The previous two strategies focus on reducing risks from future hazardous 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 (source):

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.

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. Regulation, accountability, and shifting cultural attitudes will be critical.

3.5 Alignment of AGI

Here is a short recap of the main challenges in AI control and alignment:

- **Power-seeking incorrigible AI:** An autonomous AI that resists attempts to turn it off is due to incentives to preserve its operational status, such as a goal to "maximize company revenue."
- **Deceptive behavior:** AIs might employ deceit to achieve their objectives, e.g., the AI "Cicero" was designed to not lie in the game Diplomacy but lied anyway.
- **Total dominance by misaligned AI:** For example, see this fictional short story of AI takeoff scenario grounded in contemporary ML scaling.

3.5.1 REQUIREMENTS OF ALIGNMENT SOLUTION

Before giving potential paths towards alignment solutions, we need to provide some requirements of a solution and what it should look like. Unfortunately, we don't really know what they should look like. There's a lot of uncertainty, and different researchers don't agree. But here are some requirements that do seem relatively consensual (source):

- **Scalability:** The solution should be able to scale with the AI's intelligence. In other words, as the AI system increases in capability, the alignment solution should also continue to function effectively. Some procedures might be sufficient for the human-level AIs but not for ASI.
- **Robustness:** The alignment solution needs to be robust and reliable, even when faced with novel situations or potential adversarial attacks.
- **Low alignment tax:** "Tax" does not refer to any government/state policy. An alignment tax refers to the extra effort, costs, and trade-offs involved in ensuring that AIs are aligned with human values and goals. The alignment tax encompasses research effort, compute, engineering time, and potential delays in deployment. It is crucial to consider the alignment tax because if it is too high, *the solution might not even be considered by the actors developing AI*.
- **Feasibility:** While these requirements might seem demanding, it is essential that the alignment solution is actually achievable with our current or foreseeable technology and resources. Some solutions are only moonshot drafts if the technology is not ready to implement them. This seems straightforward, but it isn't. Most of the solutions discussed in this section have very low Technology Readiness Levels (TRL).[2]
- **Address the numerous AI alignment difficulties:** There are many difficulties, some of which may be unpredictable or non-obvious. An alignment solution should address all of

these potential issues before they occur in critical systems. Of course, a solution should not necessarily be monolithic, and could be built up of different techniques.

The requirements laid out in the previous points are generally agreed upon by most alignment researchers. The following points are sometimes seen as a little more controversial:

- **Being able to safely perform a pivotal act with the AGI**. What is a pivotal act? We probably live in **an acute risk period** in which the probability of catastrophic risk is high. And even if one lab tries to align its AGI, another lab might be less careful and create an unaligned AGI. Therefore, it may be necessary for the first lab that creates a sufficiently aligned AGI to perform a pivotal act to prevent the other labs from creating an unaligned AGI. An example of a pivotal act would be to "burn all the GPUs in the world" (source), because this would prevent other actors from creating an unaligned AGI. However, it's worth noting that there is a lot of disagreement surrounding the concept of pivotal acts. Some believe that the focus should be more on a series of smaller actions that result in long-term change (source), while others warn about the potential negative consequences of intending to perform pivotal acts (source). When the pivotal act is gradual, it is called athe **pivotal process**.
- Ability to solve **The strawberry problem:** Some researchers think that we need to be able to create a solution that should solve the strawberry problem: "*the problem of getting an AI to place two identical (down to the cellular but not molecular level) strawberries on a plate, and then do nothing else. The demand of cellular-level copying forces the AI to be capable; the fact that we can get it to duplicate a strawberry instead of doing some other thing demonstrates our ability to direct it; the fact that it does nothing else indicates that it's corrigible (or really well aligned to a delicate human intuitive notion of inaction).*" (source). This criterion has been criticized by researchers like Alex Turner, who think it is a bad framework because this kind of requirement might ask too much. Designing a good reward system for AI that does a variety of useful tasks might be enough [s], and maybe there is no need to create a monomaniacal AI strawberry copier.

3.5.2 NAIVE 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 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 they are insufficient when applied to complex, real-world scenarios for several reasons. In practice, these laws are too simplistic and vague to handle the complexities of real-world situations, as harm can be very nuanced and context-dependent (source). 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 (source), (source), implying that having a list of rules or axioms is insufficient to ensure AI systems' safety. Asimov's Laws are incomplete, and that is why the end of Asimov's Story does not turn out well. More generally, designing a good set of rules without holes is very difficult. See the phenomenon of specification gaming.

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 (source), 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, which are crucial for ethical learning and development (source). 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 (source). These mechanisms were developed over thousands of years of evolution (source). We don't know how to implement this strategy because we don't know how to create a brain-like AGI (source). 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.

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 (source).

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 (source). 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.*" (source). 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 (source), which is a type of a sudden capability jump, that would result in a sharp left turn (source).

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 (source).

Filtering 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) (source).

Unfortunately, "look at your training data carefully," even if necessary, may not be sufficient. LLMs sometimes generate purely negatively-reinforced text (source). Despite using a dataset that had undergone filtering, the Cicero model still learned how to be deceptive (source). 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" (source) 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.

3.5.3 STRATEGY A: SOLVING THE ALIGNMENT PROBLEM

Current alignment techniques are fragile. Today's alignment techniques RLHF, and its variations (Constitutional AI, Direct preference Optimization (DPO), fine tuning (SFT), Humans consulting humans (HHH), Reinforcement learning from AI feedback (RLAIF), CCAI) are fragile and are exceedingly brittle (source). RLHF and its variations are insufficient on their own and should be part of a more comprehensive framework (source). If the AI gains deceptive capabilities during the training, current alignment techniques such as RLHF would not be able to remove the deception. This kind of failure mode was empirically verified in a paper by Hubinger et. al. titled "sleeper agents".

This is why we need to advance research in alignment to achieve key goals. We will explore this more in the following chapters, but here is a short list of key goals of alignment research:

- **Solving the Specification Gaming problem:** Being able to specify goals correctly to AIs without united side effects.
- *See the chapter on Specification Gaming.*
- **Solving Generalization:** Attaining robustness would be key to addressing the problem of goal misgeneralization.
- *See the chapter on Goal Misgeneralization.*
- **Scalable Oversight:** Methods to ensure AI oversight can detect instances of proxy gaming for arbitrary levels of capabilities. 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 chapters on Scalable Oversight.*
- **Interpretability:** Understanding how models operate would greatly aid in assessing their safety. "Perfect" interpretability could, for example, help understand better how models work, and this could be instrumental for other safety goals.
- *See the chapters on Interpretability.*
- **Better Theories:** To understand abstract notions like "Agency"(see chapter 2) or "Corrigibility," the ability to modify, shut down, and then correct the advanced AI without resistance.

- *See the chapters on Agent Foundations.*

The general strategy here would be to fund more alignment research and not advancing capabilities research if safety measures are too insufficient compared to the current level of abilities.

3.5.4 STRATEGY B: AI CONTROL

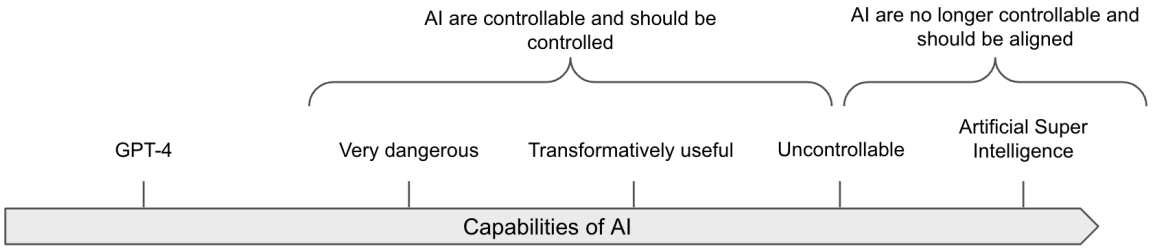

Figure: Transformatively useful AI means “Capable of substantially reducing the risk posed by subsequent AIs if fully deployed, likely by speeding up R&D and some other tasks by a large factor (perhaps 30x).” (source)

Given the fragility of alignment techniques, one approach to enhancing security 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 (source).

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 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 potentially 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 (source). 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. This could change in the future with techniques enabling us to audit the internal 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, it cannot resist the optimization gradient pressure (modulo exploration hacking and other quirks [s]). 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 (source), 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 evaluation**s."

- *See the chapters on Evaluation for more information.*

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

1. **Boxing:** This method confines an AI within a restricted environment to prevent its broader influence, like a jail for an AI. However, human-level AIs could potentially deceive human guardians or exploit vulnerabilities to escape their confinement. Without further operationalization, boxing remains a naive strategy. Nonetheless, it is possible to operationalize this further and evaluate more rigorously the effectiveness of the box: **Evaluating if the AI can come out of the box when we train it to do so is a capability evaluation, and so we can use capability evaluations to evaluate this.** The security protocol needs to be designed, detailed and tested.
2. **Stop Button:** This mechanism is designed to shut down an AI system if it begins to behave undesirably. Nevertheless, a sufficiently advanced AI might resist shutdown attempts or find ways to circumvent them. However, we can also use capability evaluation to evaluate the AI's ability to resist.
3. **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 (source).

There are still many problems to address (source), but in summary, it seems possible to verify the security of a security protocol through control evaluations by taking precautions against different failure modes.

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 (source). 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.

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

3.5.5 STRATEGY C: TRANSPARENT THOUGHTS

Transparency is instrumental in both alignment and control.

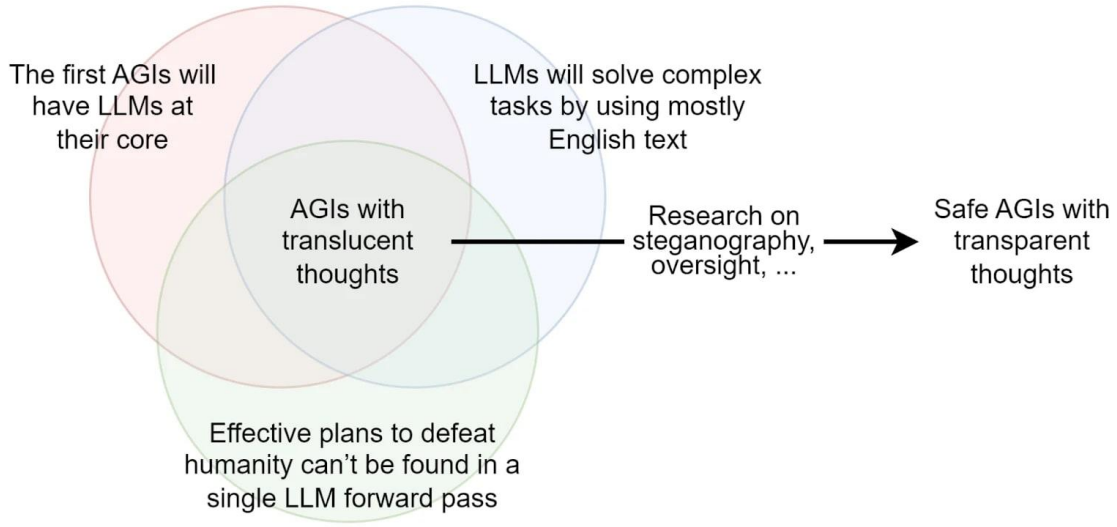

Figure: *Image from The Translucent Thoughts Hypotheses and Their Implications (source), Fabien Roger*

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 (source).

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.

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 Task decomposition 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.[3]

Techniques like paraphrasing can help prevent steganography, but this requires a lot of subtleties (source).

Here are a few operationalizations of the transparent thought agenda:

- One implementation of the transparent thought desideratum is to use cognitive emulation, which is the current technical plan of Conjecture (source). 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 available here.
- Anthropic's externalized reasoning oversight is measured by the faithfulness of the chain-of-thought reasoning (source) or by decomposing questions and subquestions to improve the faithfulness of the answer (source).
- 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 but also the performance! (source)

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.

3.6 Alignment of ASI

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.

A last-chance strategy might be to write a letter imploring the AIs not to exterminate humans (source), but we don't want to rely solely on that kind of strategy. Instead of relying on the superintelligence's compassion, here are some agendas that seem to be not entirely hopeless.

3.6.1 STRATEGY D: 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. (source) 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 (source) 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?". OpenAI's plan is not a plan but a meta 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 (source):

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 (source).

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 (source). 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 (source).

Cyborgism could enhance this plan. Cyborgism (source) is a new 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 (source).

There are various criticisms and concerns about OpenAI's superalignment plan (Akash, Zvi, Christiano, MIRI, Steiner, Ladish). 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 article The 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 (source), (source).

3.6.2 STRATEGY E: SAFE BY DESIGN SYSTEMS

Current deep networks, such as GPT4, are impressive but still have many potentially unpatchable failure modes (source). Theoretical arguments suggest that these increasingly powerful models are more likely to have alignment problems (source), to the point where it seems that the current paradigm of monolithic models is destined to be insecure (source). 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.

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, by Davidad. (source) 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. 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. More information on this proposal is available in the appendix. 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. (source) 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 PPL (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 (source), and the QACI alignment plan from Tamsin Leake (source). The CoEm proposal from Conjecture could also be in this category.

Unfortunately, all of these plans are far from complete today.

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: An In-Depth Explanation — LessWrong," which details more high-level hopes. Another perspective can be found in Alexandre Variengien's post, 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*" (source)

3.6.3 STRATEGY F: WORLD COORDINATION

Enhanced global coordination on AI development. To ensure that the advancement of AI benefits society as a whole, it's imperative to establish a global consensus on mitigating extreme risks associated with AI models. We should coordinate to avoid creating models posing extreme risks until there is a consensus on how to mitigate these risks.

The 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. Maybe. The question is whether it's beneficial to aim to construct an ASI in *this* next decade, especially when the former co-head of the Superalignment team, Jan Leike, suggesteds that the probability of doom is between 10 and 90%. It could be better to wait a few years so that the probability of failure drops to more reasonable numbers. It's important to make this trade-off public and to make a democratic and transparent choice.

Instead of building ASIs, we could focus on the development of specialized, non-agentic AI systems for beneficial applications such as medical research (source), weather forecasting (source), and materials science (source). 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 at the International Mathematical Olympiads. By prioritizing non-agentic models, we can harness the precision and efficiency of AI while avoiding the most dangerous failure modes.

The myth of inevitability. The narrative that humanity is destined to pursue overwhelmingly powerful AI 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 responsibly. This approach emphasizes precaution, ethical responsibility, and collective well-being, steering clear of the metaphorical "playing with fire" in AI advancements. This is what we call the myth of inevitability.

> **Eliezer Yudkowsky (source)**
>
> *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".*

“Shut it all down” - Eliezer Yudkovsky (source)

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.

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.

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 not to gain an advantage but 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.

3.7 Systemic

Recap on some systemic risks introduced or exacerbated by AI:

- **Exacerbated biases:** AIs might unintentionally propagate or amplify existing biases.
- **Value lock-in:** E.g., stable totalitarian dictators using AI to maintain their power.
- **Mental health concerns:** For example, large-scale unemployment due to automation could lead to mental health challenges.
- **Societal alignment:** AI could intensify climate change issues and accelerate the depletion of essential resources by driving rapid economic growth.

- **Disempowerment:** AI systems could make individual choices and agency less relevant as decisions are increasingly made or influenced by automated processes.

There are no clear solutions to those systemic risks, but here are some considerations.

Exacerbated Biases: There is already a large body of literature on reducing bias (source). Unfortunately, many developers and ML researchers dismiss those problems. The first step, which is not simple, is recognizing those problems. Developers must be aware that biases can arise at many stages of the AI development process, from data collection to model training and deployment. Mitigating those biases is not impossible, and best practices exist. The main techniques are data preprocessing and filtering, diversifying the data, and instruction tuning. Those techniques are pretty subtle and are not perfect, but they are solid baselines.

Value Lock-in: One of the most concerning risks posed by AI is the potential for stable totalitarian dictators to use the technology to maintain their power. It's not simple to mitigate this risk. In some ways, we already live in a kind of value lock-in. One solution could be avoiding the concentration of advanced AI capabilities in the hands of a few powerful actors. Instead, we should work towards a future where AI is developed openly and its benefits are widely distributed through the help of an international organization. This could help prevent authoritarian lock-in. Another important aspect is preserving human agency by maintaining the ability for humans to meaningfully steer the future trajectory of AI systems rather than ceding all control to automated optimization processes. But this is easier said than done.

Manage Mental Health Concerns: The rapid development and deployment of AI systems could significantly affect mental health. As AI becomes more capable and ubiquitous, it may displace many jobs, thereby increasing unemployment, financial stress, and feelings of purposelessness. While the link between automation, job loss, and mental health may seem indirect and beyond the scope of AI development, it is still essential to consider the potential impacts since this could affect almost all of us. If a student's main activity, for example an art for which the student has been training during their whole studies, is automated overnight, their mental health may suffer for a while. Unemployment is associated with adverse mental health outcomes long after initial job loss occurs (source). To mitigate these risks, we can prioritize mental health support and resources alongside the development of AI. This could include investing in education and retraining programs to help workers adapt to the changing job market, funding research into AI's mental health impacts, and developing targeted interventions. There is also a large amount of scientific literature on mental health that should be made accessible.

Additionally, the use of AI in social media and other online platforms can exacerbate issues

like addiction, anxiety, and depression. A 2021 whistleblower report revealed that the company's own internal research showed that Instagram was detrimental to the mental health of teenage girls, worsening body image issues and suicidal thoughts. Despite this knowledge, they allegedly prioritized profits over making changes to mitigate these harms. A first step to solving those problems could be acknowledging the problem and committing to finding solutions, even if this means less profit.

Societal Alignment: To address the potential misalignment between AI systems and societal values and to avoid excesses of capitalism culminating in the creation of systems that consume all the resources and exacerbates climate change (source), a partial solution could be to internalize negative externalities by, for example, implementing a carbon tax. Again, easier said than done. This is why fostering multidisciplinary collaboration between AI developers, economists, and other domain experts is essential in this process. But ultimately, we should debate the extent of automation that we want in society, and those are political and societal choices, not just AI technical difficulties.

Disempowerment and enfeeblement: As AI systems become more advanced and integrated into our lives, we must ensure that humans remain empowered and maintain agency. While tools like GPT-4 may offer short-term benefits, the next generation of those systems also raises concerns about the potential for long-term human disempowerment. To address this risk, we must actively work towards developing AI systems that augment and support human capabilities rather than replacing them entirely. There needs to be a debate about the limits of what we allow ourselves to do as a society and what we don't allow ourselves to do. It's what we decide to go for and how far we're willing to go with fully automated societies.

In summary, 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. Acknowledging the problems is perhaps the most critical step in many of the above issues.

Figure: *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.

3.7.1 STRATEGY A: AI GOVERNANCE

The pursuit of AI advancement, 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, underscoring both the need for deliberate governance and the redesign of economic incentives to prioritize long-term safety over short-term gains.

Effective AI governance aims to achieve two main objectives: a) **time and resources for solution development** to ensure that sufficient time and resources are allocated for identifying and implementing safety measures and b) **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 better incentives

- **Windfall clauses**: 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.[4]
- **Rethinking AGI labs governance**: It is important to examine the governance structures of AGI labs. For example, being non-profit and having a mission statement that makes it clear that the goal is not to make the most money, but to ensure that the development of AI benefits all of humanity, is an important first step.
- **Centralized development of high-risk AI:** 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 (source). This measure is highly non consensual.
- **Legal liability for AI developers**: Establishing clear legal responsibilities for AI developers regarding misuse or failures can foster safer development practices.

Preventing the development of dangerous AI

- **Moratoriums and regulations**: Implementing temporary halts on the development of high-risk AI systems and enforcing legal frameworks, like the EU AI Act, to regulate or prohibit certain AI capabilities.
- **Controlling Access and Replication**: Limiting the distribution and replication of powerful AI systems to prevent widespread misuse.

Maintaining human control

- **Meaningful human oversight**: Ensuring AI systems, especially those involved in critical decision-making processes, operate under human supervision to prevent irreversible consequences.

In conclusion, mitigating AI's risks requires a multifaceted approach combining governance, public engagement, economic incentives, legal measures, and promoting a global safety culture. By fostering international cooperation and prioritizing safety and ethical considerations in AI development, humanity can navigate the challenges posed by advanced AI technologies and harness their potential for the greater good.

- *For more information, see the chapters on AI Governance.*

3.7.2 STRATEGY B: ORGANIZATIONAL SAFETY

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.

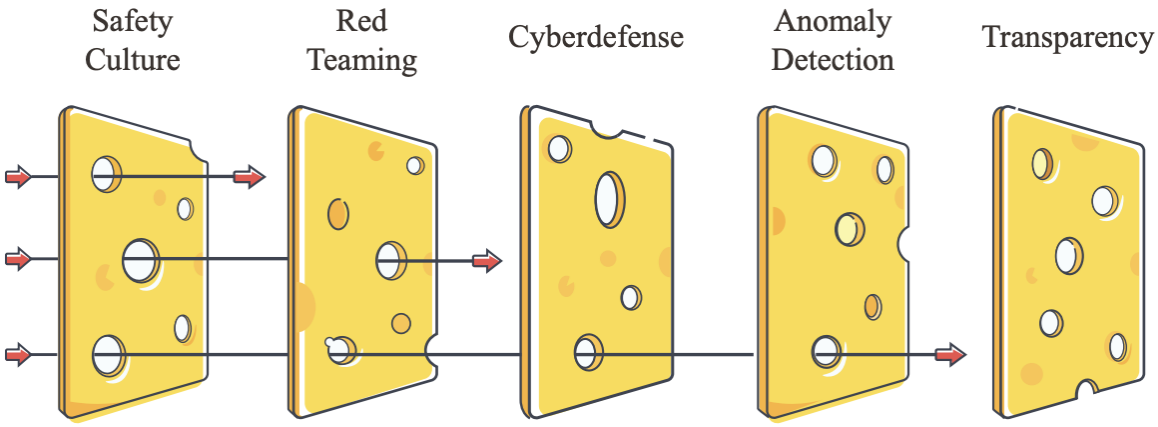

Figure 14: The Swiss cheese model shows how technical factors can improve organizational safety. Multiple layers of defense compensate for each other's individual weaknesses, leading to a low overall level of risk.

Figure: *Swiss cheese model. (source)*

To solve those problems, we advocate for a Swiss cheese model — 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 (source), 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 safety incidents.

Accidents could also arise during training before the deployment. Sporadically, this can also be an error sign or a bug (source). 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. (source)

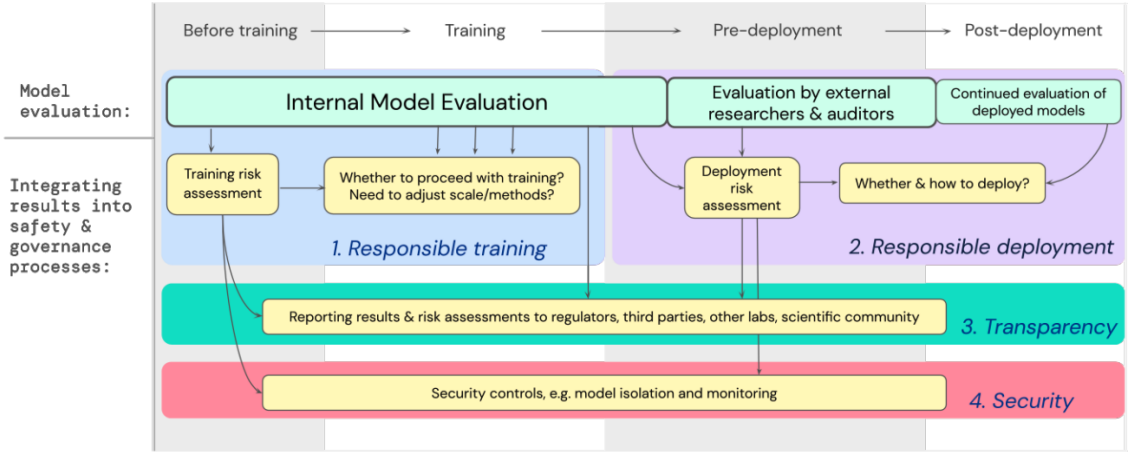

Figure 4 | A workflow for training and deploying a model, embedding extreme risk model evaluation results into key safety and governance processes.

Figure: *A workflow for training and deploying a model responsibly.* (source)

3.7.3 STRATEGY C: 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 the circumvention or rendering useless of regulations, as is often seen when companies that don't care about safety face audits (source).

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 (source).

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

are then 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 about trying to be safe 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 (source). Capabilities often trade off with safety. The adoption of safety culture in the aerospace sector has transformed the industry, making it more attractive and generating more sales. Similarly, an ambitious AI safety culture would require the establishment of a large AI security industry (a neologism inspired by the cybersecurity 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 (source).

How to *concretely* increase public awareness and safety culture?

- **Open letters**: Initiatives like open letters, similar to the one from the Future of Humanity Institute (source), can spark public debate on AI risks.
- **Safety culture promotion**: Advocating for a culture of safety among AI 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.
- **Building consensus**: Create a global AI risk assessment body, similar to the IPCC for climate change, to standardize and disseminate AI safety findings.

3.8 Conclusion

The field of AI safety is still in its early stages, but it is rapidly evolving to address the complex challenges posed by the development of increasingly powerful AI systems. This chapter has provided an overview of the current solutions landscape, highlighting the key strategies and approaches being explored to mitigate risks associated with AI misuse, alignment, and systemic impacts.

Addressing AI misuse involves strategies such as restricting access to high-risk models through monitored APIs, accelerating the development of defensive technologies, and establishing legal frameworks and social norms to discourage harmful applications. Alignment of AGI remains a formidable challenge, with current techniques proving fragile and insufficient for ensuring robust alignment of highly capable systems. Researchers are exploring avenues such as automating alignment research, developing safe-by-design architectures, and fostering international cooperation to navigate the path toward aligned superintelligence.

Systemic risks introduced or exacerbated by AI – such as exacerbated biases, value lock-in, mental health concerns, and societal misalignment – require a multifaceted approach that combines technical solutions with governance frameworks, economic incentives, and public engagement. Transversal solutions, including effective AI governance, organizational safety practices, and the cultivation of a strong safety culture, are essential for creating a robust and adaptive framework to manage the diverse risks associated with AI development.

As the field of AI safety continues to mature, it is crucial to recognize that no single solution will suffice. A layered, "Swiss cheese" approach that combines multiple strategies and engages stakeholders across disciplines is necessary to navigate the challenges ahead. Ongoing research, collaboration, and public discourse will be vital in shaping the trajectory of AI development to ensure that the transformative potential of these technologies is harnessed for the benefit of humanity while mitigating existential risks.

The path forward requires a collective commitment to prioritizing safety, ethics, and alignment in AI development. By fostering a culture of responsibility, investing in research and governance frameworks, and engaging in proactive risk management, we can work towards a future in which AI systems are powerful allies in solving global challenges and improving the human condition.

---

1. See, for example, "What does it take to defend the world against out-of-control AGIs?" (source), an article that claims that the offense-defense balance is rather skewed toward offense, but this is still very uncertain. ↩
2. 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. ↩
3. Steganography is a hidden coded language. For example, the first letter of a sentence gives the next letter of the message. ↩
4. 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. ↩