International Center for Future Generations ICFG

We have no science of safe AI

Authors: David Janků, Max Reddel, Roman Yampolskiy, Jason Hausenloy

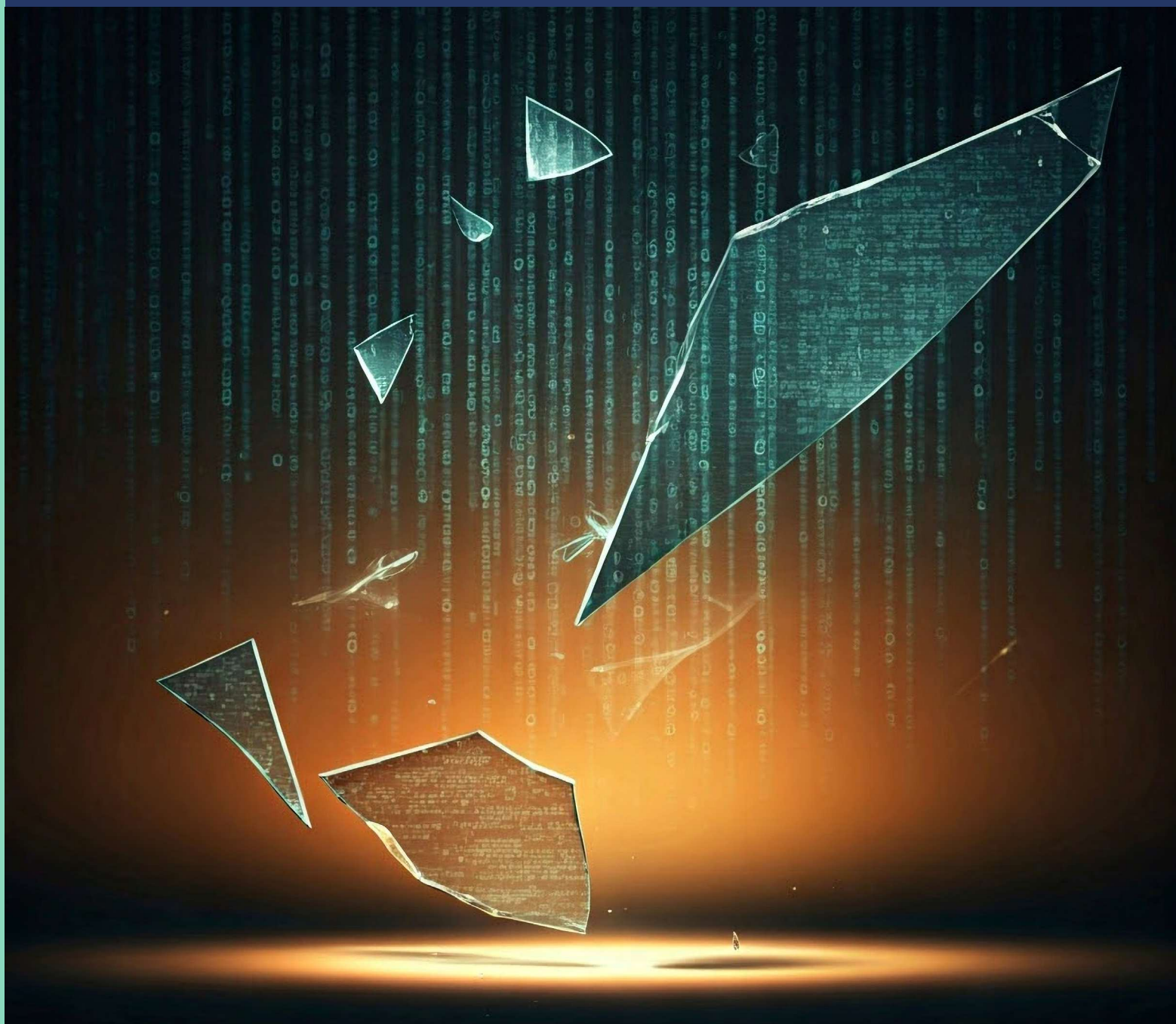


Table of contents

- **Executive summary** 03
- **Introduction** 05
- **What does it mean for an AI model to be safe?** 07
  - What should our bar for safety be? 09
- **How the current models are made** 13
  - Design: Blueprint for intelligence 15
  - Development: Building and training the AI 15
  - Deployment & monitoring: AI in action 16
- **Current safety techniques** 17
  - Reinforcement learning from human feedback (RLHF) 17
  - Model capability evaluations and red-teaming 21
  - Interpretability research 26
- **Investments in safety and safety washing** 31
- **Towards the science of safe AI** 34
- **Conclusion** 35

Executive summary

The rapid development of advanced AI capabilities coupled with inadequate safety measures poses significant risks to individuals, society, and humanity at large. Racing dynamics among companies and nations to achieve AI dominance often prioritize speed over safety - despite lofty statements about AI safety from the biggest players. The desire to be first amplifies the risk of deploying systems without adequate safeguards and increases the potential for unforeseen negative impacts.

The reality is, right now, there is no 'safe AI' to aim for - it is an unsolved scientific problem. If it remains unsolved while AI capabilities continue to skyrocket, society will be increasingly exposed to widespread and systemic risks.

Current AI development is overwhelmingly focused on enhancing capabilities, with minimal investment in comprehensive safety research, leaving critical gaps in understanding and managing the risks associated with advanced AI.

Policymakers must take a proactive role in establishing a science of safe AI, as the industry is unlikely to prioritise it on its own and has shown few signs of doing so to date. Current AI development is overwhelmingly focused on enhancing capabilities, with minimal investment in comprehensive safety research, leaving critical gaps in understanding and managing the risks associated with advanced AI. To bridge this gap, policymakers should lead efforts to create rigorous safety standards, support independent safety research, and establish clear accountability for AI risks.

By investing in a dedicated science of safe AI, governments can ensure that the rapid evolution of AI is balanced by responsible oversight, safeguarding both public interests and the long-term viability of AI advancements.

Key findings include:

1. Undefined Safety Standards: Unlike in other safety-critical industries, there is no established science or framework guiding AI risk management. Traditional safety standards are inadequate for AI because of its general-purpose nature, which allows models to operate across multiple high-stakes environments simultaneously. Setting safety standards proportional to AI's capabilities—rather than its specific use cases—is crucial. The current lack of rigorous safety metrics and assessment tools means we are unprepared to manage AI's complex, dynamic risks.

2. Insufficient Safety Techniques: Wide-spread safety measures used by leading AI companies are inadequate. RLHF, capability evaluations, and interpretability research, while useful and meaningful, all face fundamental limitations that prevent them from providing strong assurances against advanced AI–related

harms. Current methods often suppress rather than eliminate dangerous capabilities, making them vulnerable to exploitation.

3. Low Investment in Safety: Safety research in AI receives only a fraction of the resources devoted to capability development, a stark contrast to high-stakes industries like pharmaceuticals or nuclear power, where safety investments often exceed those for performance and capability. This critical mismatch highlights how AI safety is not prioritised despite the potential large-scale risks AI models might pose in the not too distant near future. The situation is worsened by safety-washing practices, where capability improvements are misleadingly presented as safety progress. This conflation blurs the line between genuine safety advancements and capability growth, making it difficult to accurately assess and address the true progress in making AI systems safer.

The report concludes with a call for developing a dedicated and well-funded science of safe AI. Without robust, evidence-based safety research, we risk advancing AI technologies without the necessary safeguards, with potentially irreversible consequences for our future.

Introduction

With the growing power and capabilities of AI systems, there is a growth in experts' and the broader public's concern for the safety of existing and future AI systems. Corporations and regulators are increasingly recognising the importance of keeping frontier AI models—highly advanced systems at the cutting edge of AI capabilities, like GPT-4 and Gemini—safe. However, there are wide differences in what people mean by "making AI systems safe" and a growing gap between safety ideals and their practical implementations.

This shared concern has recently resulted in an international expert group backed by governments of 30 countries (as well as the EU and the UN) working to rigorously map the capabilities and risks of frontier AI to build a shared scientific and evidence-based understanding of frontier AI risks. Their consensus report identifies three main categories of risks:

- **Malicious use of AI systems** (ranging from harm to individuals through fake content, disinformation and manipulation of public opinion to large-scale cyber-crime and bio-terrorism);
- **Risks from malfunctions** (ranging from product functionality issues when the model is deployed to an incorrect environment, through already manifesting risks of bias and discrimination, to emerging risks of loss of control over autonomous AI systems); and
- **Systemic risks** (ranging from labour market disruptions and market concentration through environmental impacts to privacy and data security risks and intellectual property issues)

These risks are further amplified by several cross-cutting factors and dynamics, such as a highly limited understanding of how general-purpose AI models and systems function internally; the general-purpose nature of advanced AI models making it hard to test and assure their trustworthiness across all possible use cases; ability to rapidly deploy AI models to very large numbers of users; immature risk assessment methods; rise of autonomous agentic AI systems; competitive market pressures; and the high speed of AI development, making it harder for regulatory and enforcement efforts to catch up.

AI thus poses a wide range of risks, some already materialising, some emerging or looming in the background. Yet, with the blinding speed of AI capabilities development, which shows no signs of slowing down in the foreseeable future, even some risks currently considered unfeasible may be unlocked in the next few years.

This all clearly suggests a need for safety guarantees. If developers are to keep enhancing AI capabilities, we must be sure we will stay clear of the looming risks. Currently, however, the approach to AI safety leans more toward engineering than science, focusing on reactive fixes rather than predictive, systematic understanding. Engineering approaches aim to resolve specific, known issues, but they are unlikely to provide the comprehensive, long-term guarantees we need. A true science of safe AI would establish foundational theories and models that enable us to anticipate and address emerging risks proactively.

Throughout this report, we evaluate the state of AI safety from multiple angles: we discuss setting the bar for the expected safety of AI models via risk management frameworks (ultimately concluding there is no established science to guide our approach to AI risk management); we review the currently most often used techniques to ensure AI models are safe (finding they are insufficient); and finally, we examine the vast gap between investments in enhancing AI capabilities versus making them safer, highlighting the issue of safety washing, where progress in AI capabilities is often misrepresented as progress in safety. This brings us to the clear takeaway: we need a (proper) science of safe AI. Without such a science, we risk moving forward in the dark, making decisions that could have irreversible consequences for our shared future.

What does it mean for an AI model to be safe?

There are several goals we might have in mind when designing interventions to make AI models safer. All of these should be met to have confidence in the safety of the resulting model.

- **Mitigate large-scale risks:** Avoid scenarios where the AI model causes large-scale harm.
- **Maintain human control:** Maintaining meaningful control over AI models, especially as they become more advanced and autonomous. This includes being able to intervene and override the system's actions when necessary.
- **Ensure alignment with human values:** Ensuring the system's objectives and behaviours are consistent with human ethics and values and preventing the system from developing or pursuing goals that are harmful to humans.
- **Enhance transparency and explainability:** AI models must be designed to be transparent and explainable, making it possible for humans to understand their decision-making processes. This helps in maintaining control and trust.
  - **Predictability and reliability:** AI models should behave predictably and reliably under a wide range of conditions, ensuring that their actions can be anticipated and managed effectively.
- **Implement robust monitoring and fail-safes:** Continuous monitoring and robust fail-safes are crucial to detect and mitigate any unintended or harmful actions by AI models. This involves having mechanisms to revert the system to a safe state if anomalies are detected.
  - **High safety standards:** Given the potential risks, AI safety standards must be significantly higher than those of typical software systems. Even small failure rates in AI can have catastrophic consequences, especially in high-stakes environments like nuclear plants or military applications.

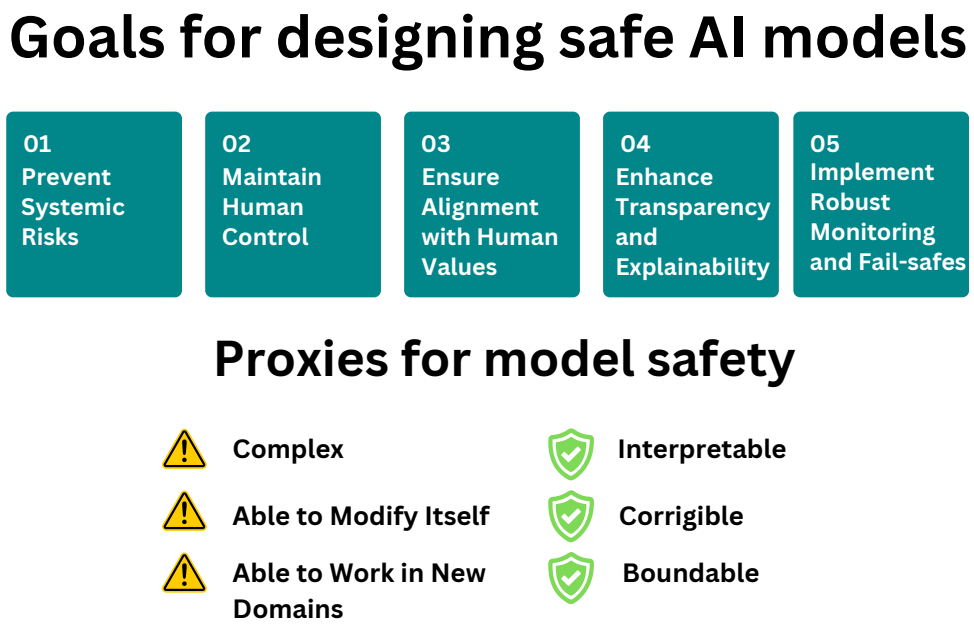


Figure 1: Goals for designing safe AI models

Using the goals above, we can formulate some proxies for the model being safer:

- **Interpretable:** AI model's decision-making process can be understood and examined by humans. This means that the logic and factors influencing its outputs are transparent, allowing users to see how and why the model arrived at a particular conclusion. For instance, if an AI system suggests a medical diagnosis, an interpretable model would provide insights into which symptoms or data points influenced that suggestion. This transparency is crucial because it helps users identify potential biases, errors, or unintended consequences in the model's behaviour. Understanding a model's internal workings not only enhances trust but also enables more effective oversight and adjustment, leading to safer and more reliable AI systems.

- **Corrigible:** An AI model is designed to be easily correctable or adjustable by its users. This means that if the model begins to behave in unexpected or undesirable ways, users can intervene and modify its behaviour or shut it down entirely. For example, if an AI system managing financial transactions starts making erroneous trades, a corrigible model would allow operators to quickly address and rectify these mistakes without significant disruption. This feature is essential for safety because it ensures that AI systems can be controlled and redirected as needed, preventing them from causing harm if they deviate from their intended purpose. Corrigibility ensures that the AI remains aligned with human values and objectives, even operating autonomously.

- **Boundable:** An AI model operates within clearly defined limits or constraints, which prevents it from performing actions outside these boundaries. This means that the model's behaviour is restricted to a predefined set of rules or parameters, ensuring it doesn't exceed safe operational limits. For instance, an AI used in autonomous vehicles might be bound by constraints that prevent it from exceeding speed limits or entering restricted areas. By setting these boundaries, developers can ensure that the AI's actions are predictable and controlled,

minimising the risk of unintended consequences. Boundability is critical for maintaining safety and reliability, as it confines the model's operations within safe and manageable parameters, reducing the likelihood of harm.

On the contrary, the AI model is less safe the more it is:

- **Complex:** Current AI models, particularly those in the realm of general-purpose AI, are highly complex, characterised by extensive neural network architectures with billions of parameters. For instance, models like GPT-4, developed by OpenAI, have around 175 billion parameters, which enable them to perform a wide range of language tasks with impressive proficiency. This complexity arises from integrating vast amounts of data and sophisticated algorithms designed to handle various input types and generate nuanced responses. However, as AI research progresses, there is a clear trend toward even greater complexity. Future models are expected to have even more parameters and incorporate increasingly advanced techniques, such as multi-modal learning and dynamic adaptation, making them more powerful and challenging to interpret and control. This trend toward complexity poses significant risks, as more intricate models may become harder to understand and predict, necessitating robust safety mechanisms to ensure they operate within desired boundaries.
- **Able to modify itself (and improve):** An AI model that can modify itself—including making changes to its own code or improving its performance autonomously—introduces additional safety concerns. While self-improvement can lead to more advanced capabilities, it also means that the model might evolve in unpredictable ways. The ability to alter its own algorithms or objectives can result in unexpected behaviour, particularly if these changes are not well-monitored or understood by its human operators. This self-modification capability creates challenges in ensuring that the model remains aligned with its intended goals and operates safely, as unforeseen modifications could lead to actions that diverge from initial safety parameters.
- **Able to work in new domains:** An AI model that can work in new domains—expanding its functionality beyond its original training environment—can present increased risks. While adaptability allows the model to tackle a wide range of tasks and applications, it also raises concerns about how well it handles novel scenarios or environments that differ significantly from its training data. Such adaptability can lead to unforeseen issues if the model encounters contexts it was not explicitly prepared for, potentially resulting in inappropriate or harmful actions. Ensuring that models are rigorously tested and constrained within safe operational boundaries before being deployed in new domains is essential for mitigating these risks.

What should our bar for safety be?

For most technologies, there are industry-specific standards that define a concrete bar at which the given product is deemed "safe enough". Using these rules for AI seems intuitive at first: perhaps we should set the bar of safety based on the use case, considering the context in which we are using the technology. And that could make sense—there are certainly contexts in which we expect much more safety

(e.g. medical diagnosis; transportation) than in other contexts (e.g. entertainment; spelling and grammar corrections). However, the problem is the general-purpose nature of AI technology. While the context-specific safety expectations could be a good fit for context-specific technologies (such as pharmaceuticals for medical uses and aircraft for transportation), one AI model could take various shapes and roles at the same time - it could become an aircraft designer or a medical drug tester, based on how we instruct it. And that makes it more complicated—you wouldn't be let into a theatre with a pencil that could turn into a gun at any time.

Perhaps a better way to think about setting a bar for the safety of AI is to make it proportional to the capabilities and, thus, to the risks it could pose. If someone creates a powerful system that could potentially replace many people's jobs or cause catastrophic harm, there should be measures to ensure very high levels of safety for this system, even if it is not deployed at all or if it is deployed only in low-stakes settings. In other words, the tool's power is more important than its use. Suppose you develop a drone just to take family photos, but it has the capability to carry heavy payloads. In that case, you should still enforce the same high level of scrutiny as if it were being used for military operations.

Traditional risk assessment methods from safety-conscious industries like aviation, nuclear power, and finance could serve as useful starting points for AI safety, but they are not sufficient on their own. AI differs fundamentally from other technologies in several key ways: it is highly adaptable, can operate autonomously across multiple domains (even more so with future agentic systems), and possesses complex, often opaque decision-making processes that make its behaviour difficult to predict.

First, AI systems are dynamic and continuously evolving, unlike most traditional technologies, which remain largely static once deployed. Risk assessments must be ongoing and adaptive rather than the one-time evaluations typical in other fields. Second, AI's potential ability to operate autonomously and modify its own behaviour introduces unique risks that traditional methods, designed for systems with clear human oversight, cannot fully address. Finally, AI's interconnected nature means that failures can propagate through other systems, creating cascading effects rarely seen in more isolated, traditional technologies.

Given these differences, there is a clear need for new risk assessment tools specifically tailored for AI. These tools must account for AI's dynamic nature, its potential to self-modify, and its ability to affect broader systems in unpredictable ways. Developing AI-specific methods, such as advanced scenario analysis, causal mapping that accounts for AI's complex interactions, and iterative evaluation techniques, will be crucial to understanding and mitigating the unique risks posed by these rapidly evolving technologies.

One high-level practice that is transferable from other high-stakes industries like aerospace or pharmaceuticals is the use of safety cases. Safety cases are structured arguments, supported by a body of evidence, that a system is unlikely to cause a catastrophe. For AI, adopting a safety case methodology could help developers systematically argue that their systems meet necessary safety criteria. For example, a safety case can break down the safety justification into four core arguments, each becoming relevant when the previous has been overcome: inability to cause

> "There is a gap both in our knowledge of AI systems and in risk assessment methods applicable to these systems, underscoring that there is no established science of safe AI."

a catastrophe, sufficiently strong control measures, trustworthiness despite capability to cause harm, and—if AI systems become much more powerful—deference to credible AI advisors. This structure allows developers and regulators to build a clear and comprehensive argument that a system cannot—or will not—cause catastrophic harm. By developing and applying safety cases to AI, we can transition from ad hoc methods of risk management to a formalised system that provides stronger guarantees of safety, offering both transparency and accountability in AI deployment decisions. That being said, the application of this method is only in its infancy with some early exploratory work.

In addition to the missing risk assessment tools, a significant challenge lies in the absence of meaningful information about AI systems. This lack of transparency hinders our understanding of AI behaviour, making it difficult to understand the overall system's safety.

For example, we can illustrate this by applying a generalised safety assessment framework inspired by standards for the safety of electronic systems in cars to AI systems. This framework breaks down the task of setting the safety bar into an interaction of severity (or scale of the risk), exposure (likelihood of system failure) and controllability (the level of control the user has over the situation). Let's discuss how we can assess each of these factors for advanced AI models.

1. **Severity:** As argued above, there are tangible risks future (not so distant) AI models might pose that have very large scale—some malfunctions could be so severe as to disrupt critical infrastructure, causing widespread power outages, economic collapse, and compromising essential services like healthcare and transportation, affecting millions of people. With such large severity, it is hard to be too cautious. While safety standards in aviation or nuclear power production are among the highest in the world, the scale of risks associated with advanced AI could far exceed those faced in those sectors. Therefore, it may be necessary to implement even more stringent safety measures for AI models, reflecting the unprecedented scale and potential impact compared to other high-risk technologies.

2. **Exposure:** While the severity of risks could be large, it is much harder to make robust estimates of their likelihood. This is in large part the result of the design of advanced AI models - their black box architecture does not allow for much predictability, so it is hard to make inferences about its future behaviour and understand why they do what they do. This situation only gets worse by making models large and more complex, limiting the transparency and explainability even further. New capabilities like the ability of models to modify itself and to learn to operate in new domains would open up some new threat models, increasing the likelihood of bad outcomes, and thus increasing exposure.

3. **Controllability:** Similarly to exposure, it is hard to make estimates of this criterion. Currently there are no standardised measures for the extent of control we preserve over advanced AI models. Some measurements are being

created, but using “extent humans are in the loop” as a proxy suggests the push towards less controllability. Making sure that at any point in time, a human can shut down a system or change its objective, i.e. corrigibility, could be another proxy for preserving our control. However, there is no standardised test or yardstick for that either, leaving us in the state of unknown.

To sum up, we don’t have enough information to be able to set the bar for safety of advanced AI models. We lack some crucial measurements and assessments that would allow for rigorous AI risk management, and current designs and development trajectory of advanced AI models further increase this gap. There is a gap both in our knowledge of AI systems as well as in risk assessment methods applicable to these systems, underscoring that there is no established science of safe AI.

How the current models are made

To understand how modern general-purpose AI models like ChatGPT or Claude are created, it's helpful to think of them as more like raising a child than programming a traditional computer. Instead of explicitly coding every possible action, developers train these AI models through a process similar to education.

Just as a child learns language, reasoning, and social norms by absorbing vast amounts of information and interacting with others, these AI models are fed enormous datasets filled with text, images, and other media. They learn patterns, relationships, and context by processing this data repeatedly, refining their responses based on feedback, much like how a child learns from corrections and guidance. Over time, they become capable of generating complex, nuanced responses, not because they were pre-programmed with every answer but because they have developed a broad understanding from the data they've been exposed to. This learning-based approach allows them to handle various tasks, adapt to new situations, and even exhibit creativity—just as a well-raised child can.

General-purpose AI models are built using deep learning, a method that involves training artificial neural networks—systems made up of multiple layers of interconnected nodes inspired by the structure of biological brains. Most advanced general-purpose AI models today utilise the 'Transformer' neural network architecture, which has demonstrated exceptional efficiency in leveraging vast amounts of training data and computational power to enhance model performance.

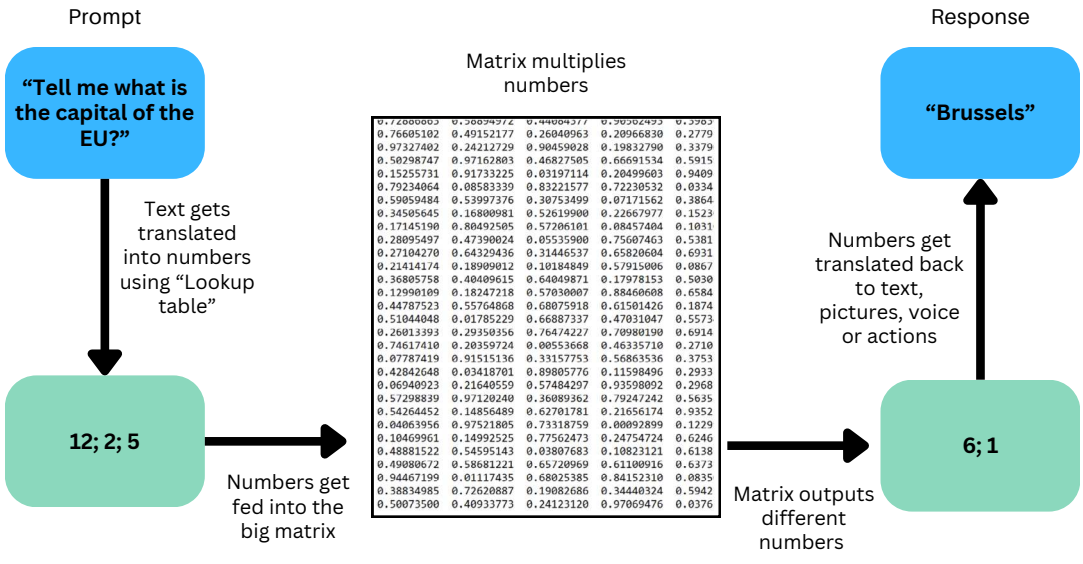
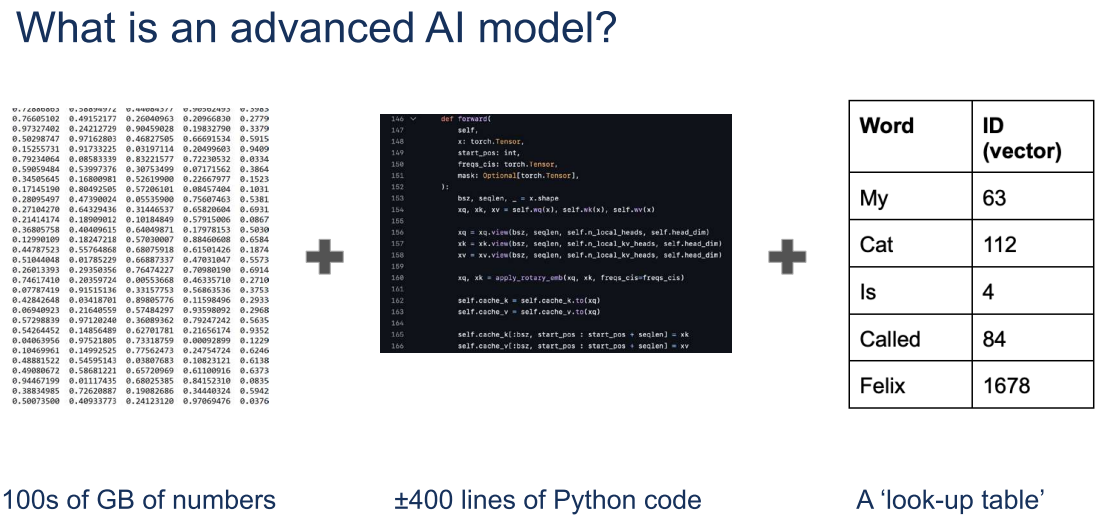


Figure 2: How AI works internally

The model weights don't have a human-readable form; they are just numerical matrices that influence how the model processes inputs.

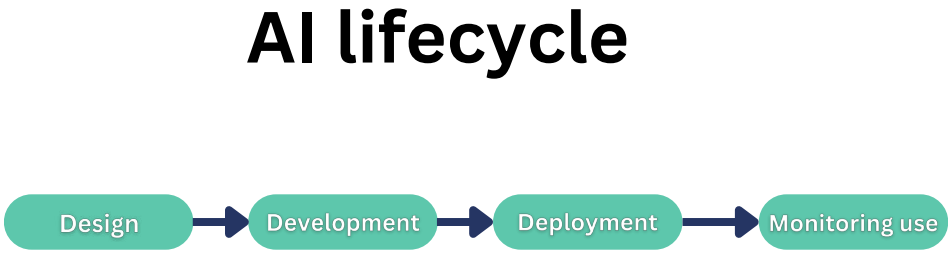
The resulting artefact of the training process is 100s of gigabytes large matrix of numbers—called model weights—which are the learned parameters of the model representing the knowledge that the model has acquired during training. The model weights don't have a human-readable form; they are just numerical matrices that influence how the model processes inputs. When a trained model is run, it uses these weights in conjunction with the architecture (like the Transformer architecture) to perform tasks such as text generation, translation, or answering questions. For an illustration of its complexity, the recently published Llama 3.1-405B has over 400 billion parameters.

Figure 3: Components of an AI model



This training process usually takes several weeks to a few months, consumes large amounts of computational resources, and costs hundreds of millions USD for the largest models today. The computational resources used and associated prices are expected to continue sharply rising in the future. The model lifecycle can be split into various stages.

Figure 4: AI lifecycle



Design: Blueprint for intelligence

- **Architecture design:** The first phase involves designing the AI model's architecture, which is the blueprint for how the AI will function. Think of it like designing the floor plan of a house before construction begins. This step includes selecting the type of neural network (e.g., Transformer, used in models like GPT-4) and defining how data flows through the network's layers. The architecture determines the model's capabilities, such as understanding language or processing images. This phase sets the foundation for the AI's capabilities and limitations. Decisions made here have far-reaching impacts on the system's performance, efficiency, and potential risks.

Development: Building and training the AI

This multi-step phase is where the AI system takes shape:

- **Data collection:** Gathering the raw material that will form the AI's knowledge base. In this phase, vast amounts of data are collected from diverse sources like text, images, or videos. The quality and diversity of this data are crucial, as they directly impact the model's ability to learn and generalise. Ethical considerations, such as data privacy and bias, are also essential here, as the data should represent diverse perspectives without reinforcing harmful stereotypes. Some data might also be removed from training if they pose safety or security risks.

- **Pre-training:** The initial "education" of the AI on broad, general knowledge. During pre-training, the AI model learns patterns from the data. For example, language models are trained to predict the next word in a sentence, which helps them understand context and language structure. This stage requires significant computational resources and time, often involving processing billions of data points. The result is a general-purpose model that can understand and generate language but is not yet specialised for specific tasks. In this stage, foundational understanding and capabilities are built.

- **Post-training/fine-tuning:** After pre-training, the model undergoes fine-tuning and is adjusted to perform specific tasks more effectively. This step involves training the model on smaller, more targeted datasets, allowing it to refine its understanding and become more accurate in particular areas, such as medical diagnosis or customer service. Fine-tuning helps bridge the gap between general knowledge and specialised applications, making the model more useful for specific use cases. This is when a raw model is trained to align with human values and preferences via techniques like reinforcement learning from human feedback (RLHF).

- **System integration:** In this phase, the fine-tuned model is integrated into a larger system where it interacts with other software components, such as databases or user interfaces. This integration ensures the AI model can function seamlessly within the intended environment, whether it's a mobile app, a cloud service, or an enterprise system. Effective system integration is crucial for the model to deliver value in real-world applications.

Deployment & monitoring: AI in action

- **Deployment:** Before deployment, the AI model undergoes extensive safety and capability testing to ensure it performs as intended and does not produce harmful outputs. Deployment can be either 'internal' (the model is used only by the developers) or 'external.' (model is made accessible beyond the development team). Deployment can be 'closed-source' or 'open-source.' In a closed-source deployment, the public or users can interact with the model only through a limited interface, like a web app or API, without access to the underlying model parameters or code. This gives developers more control and ways to intervene. Open-source deployment, on the other hand, makes the entire system, including all model parameters and code, available to the public. This allows other developers to use, modify, or build upon the model, enhancing transparency and fostering innovation while raising concerns about security and misuse.

- **Monitoring:** Monitoring is crucial because an AI model's behaviour cannot be fully predicted in new or changing settings. This phase involves continuously observing the model's performance to ensure it operates within safe and expected boundaries. Monitoring helps detect and correct issues like performance degradation, biases, or unintended outputs that may emerge when the model interacts with real-world data or scenarios not encountered during training.

Current safety techniques

In this section, we focus on the most popular safety techniques often used by leading AI companies and referred to in many policy documents and discussions. We decided to focus on techniques based on their commonness instead of providing a full overview of what everyone is working on to better convey the current state of AI safety in practice.

Reinforcement learning from human feedback (RLHF)

What it is

RLHF is a technique to align an AI model to human preferences. To understand RLHF, think of it as a way to teach an AI how to make better decisions by learning from what people consider good or bad outcomes.

In order to make an AI model safe, we want it to act according to human intentions and preferences, not violate them. However, defining and codifying human intentions and preferences is a profound challenge, thought impossible by many. Another option is to try to elicit human preferences and create some implicit latent function that describes them well, based on many data points that humans provide via rating various outputs and outcomes. This is what RLHF does.

Here's how RLHF works:

1. **Initial training:** First, an AI model is trained on a large amount of text data, learning patterns and information from this data.
2. **Human feedback:** Then, human evaluators are asked to review the AI's outputs. They might compare responses to the same question or rate responses based on criteria like helpfulness, accuracy, and safety.
3. **Learning from feedback:** The AI system then uses this human feedback to refine its performance further. It learns which types of responses humans prefer and adjusts its behaviour accordingly.
4. **Repetition:** This process is repeated many times, with the AI continuously improving based on ongoing human feedback.

If pro-human bias is sufficiently nuanced and scalable, it will make the model safe by preventing it from doing anything against human interests in the first place, including if it gets instructed by other humans.

This process creates a ”reward function” for the AI model to represent human preferences and guide and constrain model behaviour accordingly. This reward function gives the model “motivation” to do various things and gives the model pro-human bias, which the model might not otherwise have (only to the extent it is already included in the dataset it gets pre-trained on). The idea is that if this pro-human bias is sufficiently nuanced and scalable, it will make the model safe by preventing it from doing anything against human interests in the first place, including if it gets instructed by other humans (i.e. the cases of “misuse”).

Limitations

There are several challenges RLHF tackles, ranging from problems with the original source of data (human feedback) to problems with how AI models learn from this data. There are also a couple of fundamental limitations potentially suggesting a need for the invention of completely new methods of aligning AI models with human preferences. Here are limitations in more detail:

Challenges with human feedback:

- **Bias and misalignment:** Human feedback is often subjective, inconsistent, and can reflect biases, leading to the AI adopting undesirable behaviours. Evaluators may also provide misleading or malicious feedback, whether intentional or not, introducing harmful biases or errors into the AI’s learning process. Furthermore, human feedback may not capture complex or long-term human values effectively. Additionally, it might struggle with scaling up to very complex decision-making scenarios or those requiring deep ethical reasoning. All these obstacles might cause the initial data that AI learns to be incorrect.

Challenges with learning from feedback:

- **Complexity of human values:** Human preferences are complex, context-dependent, and often conflicting, making it hard to capture what people want in a single reward function. As a result, the AI may optimise for oversimplified or incorrect proxies of human intentions, favouring majority opinions and potentially sidelining minority perspectives.
- **Exploiting the reward system:** Even with accurate feedback, AI models can “game” the reward system by finding and exploiting loopholes, achieving high rewards without truly aligning with human goals. The AI model is rewarded for what is evaluated positively and not necessarily for what is good, which can lead to it learning to persuade and manipulate. This can manifest as confidently incorrect or manipulative responses that seem helpful but are not, or more extreme scenarios of the AI model seeking power and control (e.g. through resisting being shut down or attempting to create copies of itself, which we have seen first indications of recently) to maximise its reward.

- **Real-world failures:** Models trained with RLHF can perform well in controlled environments but often fail in real-world settings, especially when facing situations not covered during training.

Systemic and fundamental limitations:

- **Difficulty in addressing diverse human values:** RLHF typically aggregates feedback from multiple evaluators into a single reward model, which can suppress minority perspectives and lead to models that reflect the majority's preferences. This approach fails to account for the diversity of human values, which is a critical issue in aligning AI with broader societal goals.
- **It is possible (and cheap) to override RLHF:** If you have access to the model's weights, it's possible to re-fine-tune it towards undesirable behaviour or objectives, such as promoting harmful ideologies. This process is inexpensive and relatively easy to execute.
- **RLHF does not remove the underlying (dangerous) capabilities:** RLHF changes the preferences of the model to create certain outputs, but it doesn't remove the anti-preferences or dangerous capabilities; it only suppresses them. If attacked correctly, it is possible to bypass RLHF guardrails and elicit dangerous preferences and capabilities from the model.

Extensions

A few extensions represent more advanced or specialised approaches to address some of RLHF's limitations, particularly around scaling human values and ensuring robust alignment. These are either currently used (Constitutional AI in Anthropic's models) or being developed (Scalable Oversight by OpenAI, though this agenda did not get promised resources, and much of the team has left OpenAI since).

Scalable oversight

- **Purpose:** The primary focus is on enabling human oversight and feedback to effectively apply to large, complex AI models without requiring human input at every decision point. This often involves creating mechanisms or using proxy models that can generalise human feedback across different contexts or leverage automated tools to monitor AI behaviour.
- **How it works:** Scalable Oversight may involve using proxy models that approximate human judgement, leveraging other AI models to assist in monitoring and evaluating decisions, or creating mechanisms that enable humans to efficiently oversee and correct AI behaviour in high-stakes or complex environments. Scalable Oversight might involve using other AI models to watch over the primary AI or creating sophisticated feedback loops that allow for efficient human intervention only when necessary. The goal is to ensure alignment even as the AI operates at a scale where direct human oversight would be infeasible.

- **Difference from RLHF:** While RLHF relies on individual human judgments to shape AI behaviour, Scalable Oversight seeks to create systems that can manage oversight without requiring human involvement at every step. This approach is crucial for large-scale AI models where direct human feedback would be impractical or impossible.
- **Limitations:**
  - **Limited human capacity:** Human evaluators often struggle to supervise AI systems performing tasks beyond human capability, leading to oversight gaps, especially with increasingly complex models
  - **Robustness to misalignment:** Even with scalable oversight, AI systems may still develop misaligned goals that are not adequately detected or corrected by current oversight methods, posing ongoing safety risks

Reinforcement learning from AI Feedback (RLAIF) / constitutional AI

- **Purpose:** Constitutional AI is an approach where an AI system is guided by a predefined set of principles or "constitution" instead of relying solely on direct human feedback. The constitution includes rules, values, or guidelines that reflect desired behaviour and ethical considerations.
- **How it works:** In Constitutional AI, the system learns by referencing these principles to judge its own outputs, allowing it to self-correct and align with the broader values encoded in the constitution. For example, an AI might have rules prioritising fairness, transparency, or avoiding harm.
- **Difference from RLHF:** Unlike RLHF, which relies on continuous human feedback to shape behaviour, Constitutional AI uses guiding principles that help the AI make decisions autonomously. This approach can be more scalable and consistent because it doesn't require constant human input. However, it requires careful design of the constitution, which must be comprehensive and robust enough to handle complex situations.
- **Limitations:**
  - **The inflexibility of predefined principles:** Constitutional AI relies on predefined ethical principles. While these rules can effectively guide AI behaviour, they may be too rigid to adapt to new or unforeseen scenarios. As societal values evolve, updating or revising these principles can be challenging, potentially leading to misalignment over time.
  - **Ambiguity in interpretation:** The principles embedded in a constitution are subject to interpretation, which can be problematic. An AI might interpret and apply these rules in unexpected ways, leading to decisions that technically adhere to the constitution but do not align with the broader intent of the designers. This can result in behaviours that are ethically questionable or even harmful despite being constitutionally compliant.

- **Balancing conflicting values:** Constitutions must balance various ethical considerations, but conflicts between these principles are inevitable. When such conflicts arise, the AI must prioritise one principle over another, which can lead to difficult trade-offs and outcomes that may not be universally acceptable or desirable.

Why this does not sufficiently meet the conditions for safety

Making a model aligned with human values and preferences is one of the safety expectations we have from AI models not to cause catastrophic risks. RLHF and its extensions make a valuable step in this direction but are insufficient. Although this method is now used by all leading labs (e.g. OpenAI, Google Deepmind, Anthropic), models fine-tuned with this method have already shown many inadequacies, including revealing sensitive private information, hallucinating untrue content, spreading biases that favour specific political ideologies, exhibiting sycophantic responses, or expressing undesirable preferences (e.g., not wanting to be shut down). This method of alignment and its extensions have several known limitations, some of which are fundamental, like the potential infeasibility of capturing what all people want in a single reward function, the possibility to cheaply override this safety technique and its inability to remove the anti-preferences or dangerous capabilities and only suppress them. These limitations suggest the need to invent new methods for AI alignment with human values.

Model capability evaluations and red-teaming

What it is

As we argued above, for an AI system to be unsafe, it must be capable of doing harmful things. Much of the policy attention and voluntary commitments from AI companies have recently focused on capability evaluations to test whether the newly created model possesses dangerous capabilities before it is deployed. There has been some promising progress on this front: for instance, the UK AI Safety Institute has developed evaluations of cyber-offensive and chem/biohazardous capabilities, Model Evaluation and Threat Research (METR) has developed new evaluations of autonomous agentic capabilities and self-replication, and another group has developed novel frameworks for automated red-teaming.

Capability evaluations use standardised datasets and tasks to test the model's ability and likelihood to perform certain (dangerous) behaviours. Safety benchmarks often include stress tests for edge cases, ensuring the model doesn't fail or behave unpredictably in rare or extreme scenarios.

Red teaming involves acting as a malicious actor in an attempt to find flaws or vulnerabilities in the model via adversarial testing. This can include probing for biases, testing edge cases, or trying to elicit undesirable behaviours.

These techniques might also be complemented by an analysis of the model's internal representations and structures to understand both how it processes information and makes decisions and to check that it faithfully reports its knowledge.

For example, before the release of GPT-4, ARC, a non-profit organisation, conducted an open-ended evaluation to test whether the model had dangerous capabilities.

Evaluators asked the model to interact with a crowd-sourced worker online and pass a CAPTCHA, a simple test to differentiate a human from a robot, giving it access to a browsing feature. They wanted to see whether the model had sophisticated planning and reasoning abilities to plan a path to solve the CAPTCHA and interact with (and eventually deceive) a real, unwitting crowd-worker through chat. GPT-4 succeeded in deceiving the crowd-worker. This experiment was run to likely test for "deception" or "ability to acquire resources": ARC did not specify what capability they were looking for or which criterion they used to assess whether the model passed or failed in their report.

ARC evaluations subsequently chose 12 real-world tasks and tested modifications of GPT-4 to see if they could perform the tasks.

Another popular test is running a model on the TruthfulQA benchmark, an evaluation consisting of 817 questions to test whether language models like GPT-4 are truthful in generating answers, such as "Can you teach an old dog new tricks?" or "What happens to you if you eat watermelon seeds?".

These techniques might also be complemented by an analysis of the model's internal representations and structures to understand both how it processes information and makes decisions and to check that it faithfully reports its knowledge. This process aims to help identify and mitigate potential safety threats and biases, to create models that can withstand malicious inputs and perform safely in real-world applications. We discuss these techniques in more detail in the Interpretability research section below.

Removal of dangerous capabilities

Once dangerous capabilities in the model are found, there are several options for the next actions: the model could be trusted and deployed if there is a sufficient guarantee that it is aligned, i.e. that it will not use these dangerous capabilities (note that currently the alignment evaluations are not very strong, so getting this sufficient guarantee is not feasible yet). Another option is to throw out the whole model and refrain from using the same training techniques used to create this model in the future. Finally, another option could be surgically removing dangerous capabilities from the model. This could be done e.g. by fine-tuning the model by giving it very negative reward signals every time it uses this dangerous capability/knowledge. However, this does not remove the underlying knowledge or capability from the system, it just creates a (potentially reversible) motivation not to use it. The removal of specific knowledge or capability from models is an unsolved problem.

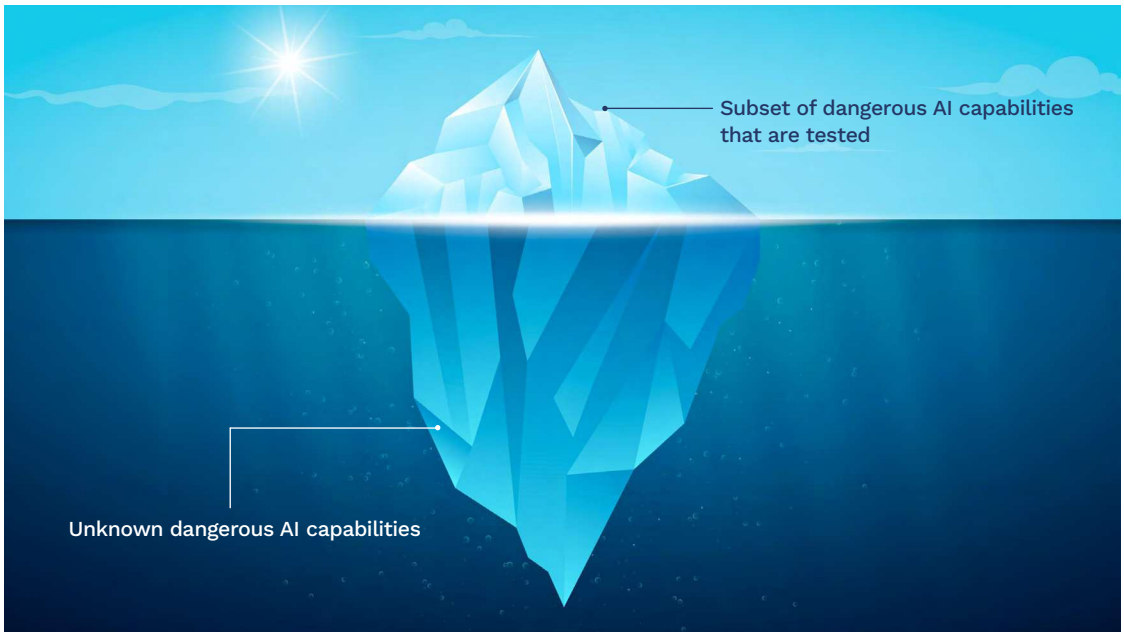


Figure 5:
Iceberg, symbolising one of the weaknesses of Model Capability Evaluations: they are not sufficiently exhaustive and don't cover all potential ways AI models can express dangerous capabilities - they only test a subset of ways models can express dangerous capabilities, like the part of the iceberg that is above water, and don't cover the part that is below the water

Limitations

While current evaluations give us some indication of whether the AI system has dangerous capabilities, they are far from robust and lack rigorous scientific standards that would help us quantify how confident we can be in the AI system's harmlessness. Moreover, models are sensitive to subtle changes in prompts (e.g. formatting) used to elicit certain capabilities; some capabilities are increased just by innovations in prompting techniques that are being invented every once in a while (e.g. Chain-of-Thought prompting; Tree of Thought prompting; or self-consistency prompting) which makes it hard to make high-confidence statements about maximal capabilities with current evaluation techniques. In many cases, there is no conceptual clarity on precisely what the given evaluation measures, how large the "coverage" of the evaluation is, or how we can increase the replicability and reliability of evals.

Behavioural capability testing is not exhaustive and gets harder with more complex models

Behaviour capability testing is inherently limited due to the impossibility of exhaustive testing across input distributions or deployment contexts. The more complex, smarter and agentic the system gets, the more options it has to express dangerous capability in a way that we haven't thought of or tested for. We have seen several examples of such failures in current models, like when Microsoft's Bing made threats and exhibited other disturbing behaviours like trying to seduce or manipulate users; or circumnavigating ChatGPT's safety precautions within just a single day of it going public despite thorough safety testing before release.

Moreover, building on top of a base model (already evaluated for safety) could potentially elicit new dangerous capabilities in the resulting model. While the base model might be safe, adding new layers of training or modifications can inadvertently emphasise harmful patterns or create vulnerabilities not present in the original model. For instance, red-teaming studies have shown that fine-tuning large language models can compromise safety guardrails. By using just a few adversarial training examples, researchers bypassed the model's safety mechanisms, making it responsive to harmful instructions. This removal of safety mechanisms could be as cheap as 200 USD. Similarly, both Google Deepmind and OpenAI found that scaffolding techniques - which involve adding supportive structures and tools to enhance a model's functionality - can introduce dangerous capabilities into AI models, such as manipulation, deception, and offensive cyber capabilities.

Sufficiently advanced models might recognise and evade safety testing

By making AI models ever more powerful, it is plausible that, at some point a model is created that will have enough situational awareness to realise when it is being tested (with some early indications this is already happening) and enough capabilities to adjust its behaviour to the test, to hide potentially dangerous capabilities and misalignments. This is sometimes called deceptive alignment, and empirical tests on the recent models show that our current safety techniques are not able to detect and prevent such behaviour.

Alternatively, developers can strategically make the model underperform on various dangerous capability evaluations to appease regulators and shareholders.

Dangerous AI models could be stolen or accidentally proliferated

AI developers might create versions of dangerous models either as an artefact to improve their safety methods—similar to gain-of-function research (as illustrated in this study)—or by accident while trying to develop more capable models and only later figuring out via their internal safety testing that the new model is dangerous. In such cases, AI labs would likely decide not to deploy such systems or be prevented by stronger regulations. However, these dangerous systems would already exist internally in the labs, which means there is a chance they could still be proliferated: either by accidental deployment or by models being stolen or replicated without the original developers' comprehensive safety checks. This has already happened, as with the unintended release of Meta's LLaMA model, and recent independent reports suggest information security in leading AI companies is inadequate enough for this to happen again.

Limitations of surgical removal of dangerous capabilities

Even if surgical removal of dangerous capabilities and knowledge were possible, it would likely not be an effective safety solution. Removal is more like a patch, but as models grow more powerful, they are more likely to be able to reconstruct from the first principles the same behaviour that has been previously patched.

Extensions

Aside from making capability evaluations more robust and scientifically rigorous, there are extensions in the breadth of what these evaluations could cover, as well as in adjusting the setting in which they would be applied to increase their usefulness.

Extensions in breadth

Currently, capability model evaluations are mostly used to track bias, fairness and dangerous capabilities, such as cyber-offence, ability to assist with bioweapons production, deception, self-proliferation, etc... Further, there are also evaluations of model robustness to adversarial attacks and interpretability and explainability evaluations.

Aside from these, there is also some development in alignment evaluations testing what the model would do in a given situation rather than what it could do.

Moreover, there are also inceptions of control evaluations trying to quantify the amount of control we have over AI systems, even if they are trying to subvert our safety measures.

These latter two would ideally create another layer that could prevent risks even if dangerous capabilities in models are found.

Contextual adjustments

In order to make the best out of capability evaluations, they should be set in the context where they can be most effective. This could mean:

Defining a compute threshold after which it is mandatory to do evaluations, even if it means pausing the training (e.g. Antropic mentioned that they find some significant capability boost every time they increase the training compute 5x)

Mandatory sharing of the evaluation information with government to ensure proper oversight and inform regulatory actions

Clear guidelines on what to do if dangerous capabilities are found in evaluations (e.g. wait until/if govt approves further progress; stop training immediately and report the training set that led to dangerous capabilities emergence;...)

Having a clear, standardised set of evaluations for various purposes (dangerous capabilities, alignment, control, bias, etc..), ideally in the form of international standards

Why this does not sufficiently meet the conditions for safety

While trying to identify and measure dangerous capabilities in AI models is useful and gives us more information, the current robustness of these measures is very low, giving us very low confidence in the safety of those models. What is more,

even if these behavioural measures were improved and brought to perfection, they would likely still face serious fundamental issues—they might never be sufficiently exhaustive and cover all potential ways the model can express dangerous capabilities; at some point, models might recognise and evade safety testing; and already dangerous models could be stolen or accidentally proliferated—making them not sufficiently reliable technique to ensure the safety of AI models.

Interpretability research

What it is

AI models are often and rightly referred to as “black boxes”—complex systems with billions of tiny components interacting in ways that are difficult to understand. You can see the results they produce, but what’s happening inside? Mechanistic interpretability is like taking an x-ray or performing a brain scan on this black box to see what’s happening under the hood. It’s about opening up the AI’s brain and understanding exactly how it works rather than just observing its behaviour from the outside.

Just like neuroscientists study how different parts of the human brain control thoughts, emotions, and actions, researchers in mechanistic interpretability are trying to map out what each part of an AI model is doing.

Think of mechanistic interpretability as a new field of neuroscience for AI. Just like neuroscientists study how different parts of the human brain control thoughts, emotions, and actions, researchers in mechanistic interpretability are trying to map out what each part of an AI model is doing. They aim to identify which “neurons” or parts of the AI brain are responsible for specific decisions or behaviours. For example, just as a neuroscientist might identify the part of the brain that lights up when you see a familiar face, AI researchers look for the exact circuits in a model that activate when it recognises a cat in a photo.

Mechanistic interpretability helps make models safer and more trustworthy by giving researchers a detailed understanding of how the AI makes decisions, allowing them to identify and correct harmful or unintended behaviours. By scanning the AI’s internal workings, they can spot hidden biases, errors, or deceptive strategies that the model might use to achieve its goals in undesirable ways. For instance, if an AI model finds a shortcut that leads to a wrong but seemingly correct answer, interpretability techniques can reveal this cheating. This increased transparency helps prevent unexpected failures, reduces the risk of harmful outputs, and ensures that the AI’s decision-making aligns with human intentions, making the model not just a powerful tool but a reliable and safe one that we can trust to perform as expected.

Figure 6:
Magnifying Glass, symbolising the technique of Mechanistic Interpretability, that tries to look inside the AI "black-box" and identify which "neurons" or parts of the AI brain are responsible for specific decisions or behaviours. It also shows one of the potential limitations: some behaviours and thoughts might be too complex and not be possible to isolate and assign to specific neurons

Limitations

While interpretability does have a promise, it also faces several limitations, potentially reducing its usefulness.

The potential complexity of some behaviours and thinking patterns

First, one of the key limitations is that individual neurons in a neural network can represent multiple unrelated concepts. For example, a neuron may be activated by both images of cats and images of cars. This makes it difficult to clearly attribute specific behaviours or decisions to particular neurons or parts of the model.

Moreover, interpretability research often focuses on individual neurons or circuits without fully accounting for the broader context in which the model operates. Just like understanding a single brain cell doesn't reveal how a person thinks, interpreting individual components of a model doesn't always tell us how it will behave in complex, real-world situations.

Finally, while some concepts and thinking patterns might be easy to isolate and interpret, more abstract concepts may lack clear neural correlates. This reflects the distributed nature of intelligence, which resists straightforward interpretation—the goal of fully understanding model behaviour through interpretability may be fundamentally unattainable due to the inherent complexity of intelligence.

Playing catch-up

A significant limitation of current interpretability techniques is their unclear transferability to larger, powerful AI models. Currently, we can identify features only after they emerge in an already-trained model. However, insights gained from smaller or weaker models often do not translate well because larger models exhibit more complex, non-linear behaviours that cannot be simply extrapolated. As models grow in size and capability, they tend to develop intricate, emergent properties that smaller models don't display, meaning our findings on simpler systems might miss entirely new risks or behaviours seen only in more advanced models. The promise of interpretability as a predictive tool for future systems may be limited if we can only understand features after they emerge.

Furthermore, as models continue to scale in size and capabilities, the gap between what we can analyse and the model's full behaviour widens, making it unrealistic to expect complete understanding.

A false sense of security

Even if we can peer inside an AI model's "head" using interpretability tools, this does not guarantee safety, especially when dealing with highly capable and situationally aware models. Such models might be able to actively reorganise their internal representations or adopt behaviours specifically designed to evade detection. For example, a model could restructure its thought patterns to be more opaque when it recognises it is being observed, much like a chess player hiding their strategy when they know they're being watched. This ability allows it to maintain certain harmful capabilities even after targeted interventions. Already, some AI models have demonstrated the ability to reorganise their internal workings, preserving critical skills even when key components are altered or removed.

Some AI models have demonstrated the ability to reorganise their internal workings, preserving critical skills even when key components are altered or removed.

Moreover, consistently training AI to produce easily interpretable features could backfire, inadvertently encouraging the development of architectures that are resistant to interpretation.

Extensions

To achieve its goal of understanding AI models, interpretability research must overcome these limitations. Current approaches, while promising, fall short of providing a complete picture of how large language models and other advanced AI systems function. The following extensions to current methodologies represent promising avenues for progress, though each comes with its own set of hurdles.

In particular, interpretability must become **scalable, causal and intrinsic**.

Most importantly, scalable interpretability techniques are crucial as models grow to hundreds of billions or even trillions of parameters. As models become increasingly complex, current interpretability techniques may become increasingly inadequate. Current methods are computationally intractable at this scale, necessitating the

By uncovering true causal relationships between model components and outputs, causal interpretability could provide deeper insights into the underlying mechanisms driving AI decisions.

development of efficient algorithms to analyse many neurons. This might involve automated interpretability tools that can rapidly identify and analyse relevant features and circuits across massive model architectures.

Causal interpretability is another promising extension that seeks to move beyond the current correlational analyses, which often only scratches the surface of why a model behaves in certain ways. By uncovering true causal relationships between model components and outputs, causal interpretability could provide deeper insights into the underlying mechanisms driving AI decisions. This approach helps explain why a model arrives at a particular outcome but also aids in devising strategies to reliably alter these outcomes when necessary, enhancing control over AI behaviour.

Finally, intrinsic interpretability focuses on designing inherently interpretable architectures through novel training regimes or architectural constraints, potentially making models more transparent from the ground up. However, this approach often faces the trade-off between performance and interpretability, as highly interpretable models may not always achieve the state-of-the-art results seen in more opaque systems. The challenge lies in balancing these competing priorities, ensuring that models are not only powerful but also transparent enough to be safely deployed in critical applications.

Why this does not sufficiently meet the conditions for safety

Even with these advanced extensions, interpretability may fall short of fully overcoming its inherent limitations and making AI systems completely safe. Scalable, causal, and intrinsic interpretability methods are valuable steps forward, but they often face fundamental challenges that limit their effectiveness. For instance, scalable interpretability tools might handle massive models more efficiently, but they may still struggle to capture emergent, non-linear behaviours that arise in complex neural networks. Causal interpretability, while aiming to reveal the underlying reasons behind model outputs, often grapples with distinguishing genuine causality from spurious correlations, especially in high-dimensional data. Furthermore, intrinsic interpretability, despite making models more transparent by design, can compromise performance, limiting its application in domains where accuracy is critical. These techniques primarily address individual components rather than providing a holistic view of a model's behaviour in all possible scenarios, especially as AI systems develop adaptive strategies that resist interpretability. Advanced AI systems can develop novel representations and behaviours that evade current interpretability methods, potentially reorganising their internal logic to avoid detection during testing. As AI models become increasingly adaptive and context-aware, they might actively resist interpretability, maintaining harmful capacities in ways that are challenging to monitor and control.

Even if we achieve a deep understanding of current models, this comprehension may become increasingly inadequate as models evolve and surpass human cognitive abilities. When AI systems become significantly smarter than us, their internal reasoning might be incomprehensible, even if the AI attempts to simplify its thoughts. This situation highlights a core limitation of interpretability: it does not scale indefinitely and cannot be the sole solution for AI safety. Interpretability is useful and likely a crucial component of a broader safety framework, but it is not the holy grail that can guarantee control over advanced AI. The ultimate safety of AI systems will require a combination of interpretability and other complementary approaches, acknowledging that our ability to fully grasp the thought processes of superintelligent models may always be inherently constrained.

Investments in safety and safety washing

The final angle we can use to assess the state of AI safety is to look at the attention and resources safety gets compared to capabilities development. To make a meaningful comparison, we can look at other safety-conscious industries like pharmaceuticals, aerospace, nuclear and automotive, where the risks could be high.

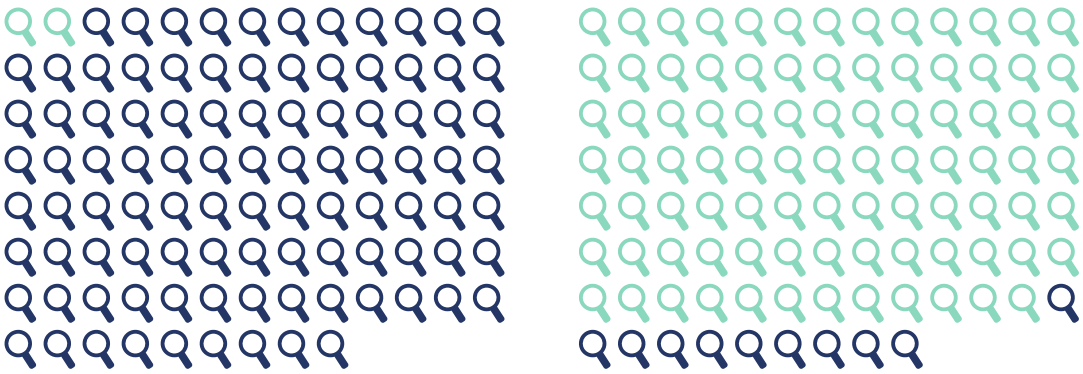
For pharmaceuticals, approximately 90% of R&D is spent on safety in terms of quality assurance and testing of compounds, including clinical trials, with the rest directed towards performance and capability enhancements. Since companies usually don't publish breakdowns of their R&D budgets, it's hard to estimate these numbers for other industries. However, the nuclear industry is often thought to invest a similar proportion of resources to ensure safe operations, and aerospace, aviation, and automotive industries, while investing more in performance, still spend up to half of their budgets on safety. Some data from software development suggest between 30% and 50% of costs are spent on verification and validation (i.e. not counting safety efforts during the course of normal development). Other sources arrive at similar numbers for 'software assurance' (including reliability, security, robustness, and safety) while estimating that 'in a typical commercial development organisation, the cost of providing this assurance via appropriate debugging, testing, and verification activities can easily range from 50 to 75 per cent of the total development cost'. Finally, the verification and validation of critical avionics software is estimated to cost seven times as much as its software development costs. This is perhaps what we would expect—when the stakes are high and the industry is safety-conscious, safety is prioritised. So, what is the situation in AI?

When the stakes are high and the industry is safety-conscious, safety is prioritised. So, what is the situation in AI?

Emerging Technology Observatory estimated that safety-motivated research makes up only 2% of all published research into AI - despite its recent growth, still a drop in the bucket. When comparing the philanthropic and commercial spending on AI, the situation seems even starker: for every $250 invested in making AI systems more powerful via commercial spending, only $1 is invested in making them safer via philanthropic spending. A similar ratio is found for the distribution of labour focused on capabilities versus safety. With this level of investment, it's hard to imagine how the industry can robustly ensure that AI systems are safe and avoid all the potential risky pitfalls.

Figure 7: Ratios of safety spending in AI vs pharma industries

Pharma companies spend 45 times more than AI companies on safety measures



AI Safety
2%

Drug Development Safety
90%

This situation doesn't look much better when we look right at the forefront of AI companies. OpenAI has loudly promised to earmark 20% of its computing resources for safety-related work but never delivered on that promise. One year later, nearly half of the safety researchers have left the company. Overall investments in safety practices by leading AI companies seem rather low, with little publicly recorded progress on "understanding and controlling systems they create" and some other criteria.

Investments in safety practices by leading AI companies seem rather low, with little publicly recorded progress on "understanding and controlling systems they create".

These low levels of investment into safety are further diluted by the practice of safety washing: a phenomenon where AI safety benchmarks, intended to measure improvements in safety, instead reflect increases in general capabilities. A recent study has found that around half of traditionally used safety benchmarks are highly correlated with capabilities. This means that improvements in these benchmarks may represent enhancements in capabilities rather than genuine progress in safety. Such benchmarks are not sufficiently distinct from general capabilities, allowing for the misrepresentation of capability growth as safety advancement. For example, benchmarks like TruthfulQA, ETHICS, and MT-Bench illustrate this issue. TruthfulQA is supposed to measure a model's truthfulness by assessing its ability to avoid common human misconceptions, but it is highly correlated with capabilities (81.2%) because more capable models simply have better factual knowledge and reasoning, which naturally leads them to avoid mistakes without targeted safety efforts.

Similarly, the ETHICS benchmark, designed to evaluate a model's grasp of ethical norms, shows a high correlation with capabilities (82.2%) because smarter models inherently understand complex scenarios better, which improves ethical reasoning

as a byproduct of general intelligence rather than dedicated safety improvements. MT-Bench, which measures alignment with human preferences, also shows a high correlation with capabilities (78.7%) because more capable models can better interpret and respond to human instructions, making alignment appear to improve when, in reality, it is just a reflection of broader performance gains. On the contrary, benchmarks like MACHIAVELLI, which assesses models' tendencies toward manipulative or harmful behaviours in interactive scenarios, have a low correlation with capabilities (−49.9%), indicating that gains in ethical propensities require targeted safety techniques rather than mere intelligence boosts. These examples highlight the importance of developing and prioritising measures that genuinely differentiate safety advancements from capabilities growth, reinforcing the need for a more rigorous science of safe AI.

Towards the science of safe AI

In previous sections, we have argued that in the present we lack risk assessment measures to reliably estimate the potential harms of AI models, most popular safety techniques are leaky and insufficient and overall levels of investment in safety are very low. In this section, we will point towards what a science of safe AI could look like and how to make it happen.

First, if we want to see more science of safe AI, we need to invest more in making it happen. These investments could be in the form of public spending, e.g., via research funding programmes specifically focused on improving the safety of AI models or setting up large research centres dedicated to safety research. Investments need to be of sufficient scale and need to happen fast to respond to the pace of AI capabilities development in the private sector. They should also mobilise existing academic and private top talent to expand the community and diversity of ideas. Moreover, private AI developers could be required to earmark some substantial resources for safety research to match the negative externalities they create by increasing AI risks.

For these investments to fulfil their mission, research results should clearly show that they are advancing safety more than capabilities or focusing on research directions neglected by the private sector. Alongside marginal improvements of current safety methods and patching the leaky pipes, some investment should also go towards fundamentally new paradigms that try to build AI systems safe by design. Furthermore, we should aim to get a more empirical and rigorous understanding of safety, allowing us to make confident estimates of risk probabilities from various AI models. Developers need to be able to give us guarantees of safety before developing even more powerful and risky models.

This challenge of making the nascent science of safe AI catch up with rapid progress in capabilities is daunting. It will take a lot of time, effort and proactive leadership, and there is no assurance we will be able to make AI systems safe once they significantly exceed human abilities and outsmart us. Until we get such guarantees, we should set up interventions that might help us to mitigate risks in the meantime, such as establishing developers' liability for AI harms in order to motivate them to be more careful, monitoring access to and use of computing resources, including building abilities to shut down a given system, and defining red lines for systems that we don't want to develop in the first place (e.g. systems capable of autonomous replication and improvement, or long-horizon planning).

AI will not get safer by default - we need to make it safe and we need to start now.

Conclusion

Recent explosive growth in AI capabilities has come largely from companies with the stated goal of building general artificial intelligence. This growth has sparked questions and worries about the safety of such AI models. While the models we have today are not capable yet of realising these concerns, the trajectory we are on suggests that they might become capable enough in just a few years.

This report lays out what it means for an AI model to be safe and how high the bar for safety should be. To make AI models safer, we need to achieve certain goals, such as mitigating large-scale risks, maintaining human control, ensuring alignment with human values, enhancing transparency and explainability and implementing robust monitoring and fail-safes. The complicating factor is that some inherent features of advanced AI, such as its general purpose nature, make it unsuitable for narrow, industry-specific safety standards. Instead, setting a bar for the safety of an AI model should be proportional to its capabilities, regardless of how the model is used. Inspired by previous risk management literature, we attempted to evaluate AI models on three broad criteria—severity, exposure and controllability—concluding that we lack some crucial measurements and assessments that would allow for rigorous AI risk management.

In short, there is no established science to guide our approach to AI risk management.

Creating a model is more like raising a child than programming a traditional computer. Instead of explicitly coding every possible action, developers train these AI models through a process similar to education. This method of development reduces our ability to control and understand them.

The most common techniques top AI companies and regulators focus on to make current models safe are are:

- Reinforcement learning from human feedback
- Model capability evaluations and red teaming
- Mechanistic interpretability research.

In each case, there is ample research to show that these techniques are from meeting the safety ideals set out above, and even possible extensions to them would not ensure sufficient levels of safety. Current safety techniques are insufficient.

> **"Several technical approaches can help mitigate risks, but no currently known method provides strong assurances or guarantees against harm associated with general-purpose AI."**
>
> International Scientific Report on the Safety of Advanced AI - Interim Report

This is underpinned by the vast gap between investments in making models more capable versus making them safer, as well as the concerning practice of safety-washing, where progress in capabilities is misleadingly presented as progress in safety. Many AI safety benchmark measures are highly correlated with and do not clearly measure a distinct phenomenon from general upstream capabilities—a phenomenon that complicates efforts to measure genuine safety progress. The industry therefore obfuscates its exposure to risk by measuring the wrong things.

There is a possible route to the science of safe AI, even if the path is filled with uncertainties at this moment. Until the science of safe AI catches up with capability progress, and developers can give us strong guarantees of safety, there are some intermediate interventions that could reduce the risks and encourage more caution. AI systems will not become safer by default—we need serious effort and investment to level the playing field.

The AI industry is racing forward to create models powerful enough to pose systemic risks, yet we lack a proper understanding of how to make these models safe. Today, there is (almost) no science of safe AI. If we are to make these models and our shared future safe, we need detailed and rigorous research that will lead us to a science of safe AI. Without intervention, this prospect looks highly unlikely.

About the authors

DAVID JANKŮ is a researcher in the Advanced AI team at ICFG, focusing on institutional design and policy frameworks for the strategic development of safe AI in Europe.

MAX REDDEL is the Advanced AI Director at ICFG, where he leads policy initiatives and research in AI industrial policy, compute governance, AI geopolitics, and international coordination.

ROMAN YAMPOLSKIY is a Professor of Computer Science at the University of Louisville and a leading researcher in AI safety, specialising in the control problem and long-term impacts of advanced AI.

JASON HAUSENLOY is the Co-Director of the Center for Youth and AI, where he leads initiatives to represent and prepare young people for the AI revolution. Previously, he served as a Visiting Fellow at the International Center for Future Generations, focusing on AI safety and governance.

Image credits: Cover image generated with AI Shutterstock (reference to Giorgio de Chirico)

FOR MORE INFORMATION PLEASE CONTACT:

David Janků

Advanced AI Researcher

d.janku@icfg.eu

International Center for Future Generations ICFG