

Safety cases for frontier AI

Marie Davidsen Buhl* Gaurav Sett Leonie Koessler

Jonas Schuett Markus Anderljung

Centre for the Governance of AI

Abstract

As frontier artificial intelligence (AI) systems become more capable, it becomes more important that developers can explain why their systems are sufficiently safe. One way to do so is via safety cases: reports that make a structured argument, supported by evidence, that a system is safe enough in a given operational context. Safety cases are already common in other safety-critical industries such as aviation and nuclear power. In this paper, we explain why they may also be a useful tool in frontier AI governance, both in industry self-regulation and government regulation. We then discuss the practicalities of safety cases, outlining how to produce a frontier AI safety case and discussing what still needs to happen before safety cases can substantially inform decisions.

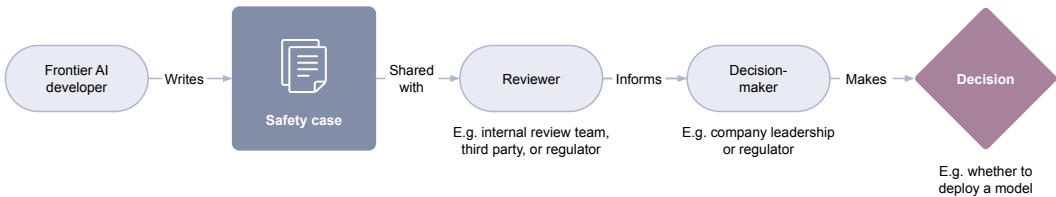


Figure 1: Process for using a safety case to inform a decision

*Corresponding author: marie.buhl@governance.ai.

Contents

1	Introduction	4
2	Safety cases and frontier AI	5
2.1	What is a safety case?	5
2.2	What is a frontier AI safety case?	6
2.3	How do safety cases relate to existing frontier AI safety frameworks?	6
3	The case for safety cases	8
3.1	Safety cases in self-regulation	8
3.2	Safety cases in regulation	9
4	Components of a safety case	12
4.1	Scope	12
4.2	Objectives	12
4.3	Arguments	14
4.4	Evidence	15
5	Implementation challenges	16
5.1	Technical challenges	16
5.2	Institutional challenges	17
6	Policy recommendations	19
7	Conclusion	20
	References	21

Executive summary

What is a safety case? (Section 2)

- A safety case is a structured argument, supported by evidence, that a system is safe enough in a given operational context. Safety cases are typically prepared by a developer ahead of a major decision (e.g. whether to deploy a system), reviewed by an independent third party (e.g. a third-party auditor), and shared with decision-makers (e.g. the board of directors or a regulator).
- Safety cases are common in other safety-critical industries such as nuclear power, aviation, and autonomous vehicles.
- Safety cases could complement existing frontier AI safety frameworks, which are organization-level policies for managing risks from frontier AI systems. By contrast, a safety case is a system-level assessment of such risks. Developers could use safety cases to explain how they have adhered to their safety framework, including its more subjective components.

Why use safety cases for frontier AI systems? (Section 3)

- Safety cases could be used internally by developers to inform high-stakes decisions (e.g. whether to deploy a system), as well as by regulators to assess compliance with safety requirements.
- A benefit of safety cases is that making an explicit, structured argument that a system is safe enough is a useful way to stress test risk assessments. In particular, it could help developers and third parties to spot shortcomings in developers' risk assessments.
- Another benefit of safety cases is their flexibility. Safety cases allow developers to assess and mitigate risks with the methods most suitable to a particular system. This makes them a future-proof tool that will remain useful even if frontier AI continues to develop rapidly. It also capitalizes on developers' expertise and resources.

What would a frontier AI safety case look like? (Section 4)

- A safety case has four key components: Objectives (that must be met for the system to be safe enough), arguments (that the objectives have been met), evidence (that the arguments are true), and scope (in which the safety case holds).
- Safety cases for current systems could be based on the arguments implicit in existing safety frameworks. Namely, they could argue that a system is not capable enough to cause catastrophic harm by evaluating if the system has certain dangerous capabilities.
- For systems with significantly more dangerous capabilities, safety cases could argue that effective safeguards are in place or that the system will not use its capabilities to cause catastrophic harm.

What challenges need to be addressed before frontier AI safety cases can inform decision-making? (Section 5)

- One challenge will be to develop methodology for frontier AI safety cases. This is still early-stage; more research and investment is needed before best practices can be established.
- Another challenge will be to develop adequate safety cases for future systems with more dangerous capabilities, as we do not yet know how to reliably assure the safety of such systems. Significant research into novel safety techniques is needed.
- A third challenge will be for developers and regulators to establish processes and build sufficient capacity to effectively review safety cases.

What should be done now? (Section 6)

- Developers should begin to produce and share safety cases; commit to using safety cases in future deployment decisions; build up internal capacity and processes for producing and reviewing safety cases; and share lessons with other actors.
- Governments should encourage companies to produce and share safety cases; conduct or support relevant research; advance a third-party ecosystem; and consider using safety cases to assess compliance with potential future regulation.

1 Introduction

Frontier AI systems – highly capable general-purpose AI systems that can perform a wide variety of tasks and match or exceed the capabilities present in the most advanced systems (DSIT, 2024)¹ – may pose severe risk to society.² Existing systems can already help conduct cyber-attacks (Fang et al., 2024a, 2024b; NCSC, 2024) and show early signs of manipulative and deceptive capabilities (OpenAI, 2024; Park et al., 2024; Scheurer et al., 2024). Future systems may be able to assist users in producing biological weapons (Mouton et al., 2023; Soice et al., 2023; Urbina et al., 2022) or act with increasing agency, making them difficult to control (Chan et al., 2023; Cohen et al., 2024; Kinniment et al., 2024). Eventually, AI systems may even cause catastrophic outcomes (Bengio et al., 2024; Grace et al., 2024; Hendrycks et al., 2023). If capabilities continue to advance as rapidly as in recent years, these risks could increasingly become a reality.

Given the potential risks, frontier AI developers³ should be able to explain why they think their systems are safe enough⁴ to develop⁵ or deploy.⁶ Safety cases are one way in which developers could produce and communicate such explanations. A safety case is a structured argument, supported by evidence, that a system is safe enough to deploy in a given way (MoD, 2007). Safety cases are used in many safety-critical industries, such as nuclear power, aviation, and autonomous vehicles (The Health Foundation, 2012; Inge, 2007; Sujan et al., 2016). There has recently been increasing interest in using them for frontier AI as well – from academics (Bengio et al., 2024; Clymer et al., 2024; Wasil et al., 2024; Yohsua et al., 2024), governments (Irving, 2024), and developers (Anthropic, 2024; Google DeepMind, 2024). However, there is still little clarity on what frontier AI safety cases would look like and how they could be used.

This paper explores the idea of using safety cases for frontier AI. Our scope is restricted in four ways. First, we focus on frontier AI systems as opposed to AI systems in general. We chose this focus because frontier AI systems are among the systems most likely to pose severe risks and present unique assurance challenges that warrant novel analysis. However, safety cases may also be a useful tool for other high-risk systems, such as biological design tools (Sandbrink, 2023) and other advanced narrow-purpose systems in high-risk domains. Indeed, safety cases are already used for autonomous vehicles (Favaro et al., 2023) and defense-related software (MoD, 2007). Second, we focus on catastrophic risks⁷ since they are the primary focus of companies’ existing safety frameworks (Anthropic, 2024; OpenAI, 2023; Google DeepMind, 2024; Magic, 2024), though safety cases could also be used to address other risks. Third, we focus on the US, UK, and EU contexts both due to their significance in the current conversation about frontier AI governance and because they are the jurisdictions we are most familiar with. Fourth, we use deployment as our primary example, though safety cases may also be useful in informing development decisions (e.g. whether to begin a training run).

The article proceeds as follows. Section 2 provides an overview of what safety cases are. Section 3 discusses the benefits and downsides of using safety cases in industry self-regulation and government regulation of frontier AI. Section 4 outlines what frontier AI safety cases could look like in more detail. Section 5 discusses potential challenges to using safety cases in frontier AI regulation. Section 6 provides recommendations for developers and policymakers. Section 7 concludes with a summary of our main contributions and suggestions for further research.

¹But note that the term “frontier AI” has been criticized (Helfrich, 2024).

²By “risks to society”, we mean risks of significant harm to large groups of people. This does not include legal, financial, or reputation risks to frontier AI companies themselves.

³By “frontier AI developers”, we mean organizations that design and train complete frontier AI systems. This currently covers a handful of private companies, but could in theory also cover non-profit or governmental organizations. It does not include downstream developers who build products based on frontier AI systems.

⁴We use “the system is safe enough” interchangeably with “the system does not pose unacceptable risk”.

⁵By “development”, we mean designing, training, fine-tuning, and applying safeguards to an AI system.

⁶By “deployment”, we mean releasing a system to certain users or applying it to certain real-world tasks. There are many possible types and degrees of deployment (Solaiman, 2023). For example, the developer may open source the model weights (Kapoor et al., 2024; Seger et al., 2023) or only allow access via an API (Shevlane, 2022). In writing this paper, we primarily had in mind internal deployment, although safety cases may also be used to inform internal deployment decisions.

⁷By “catastrophic risk”, we mean risks of extremely large-scale harm, for example damage in the tens of thousands of lives lost, hundreds of billions of dollars of economic or environmental damage, or significant adverse disruption to the social and political order (Shevlane et al., 2023).

Components	Significance
Argument ("An argument...")	A safety case must comprehensively justify why a system is safe. It must explain the relevance and sufficiency of the available evidence.
Evidence ("...supported by evidence...")	A safety case must provide support for claims and clearly state its assumptions. It must document how safety has been assessed and achieved.
Objectives ("...that a system is safe enough...")	A safety case must ultimately be about outcomes ("the system is safe enough") rather than a product (e.g. "the system design adheres to safety standards") or process (e.g. "the system was tested with SOTA techniques").
Scope ("...in a given operational context.")	A safety case must specify the conditions under which the argument is valid.

Table 1: Summary of the four key components of a safety case

2 Safety cases and frontier AI

This section outlines the concept of safety cases. We cover what safety cases are (Section 2.1), what safety cases for frontier AI might look like (Section 2.2), and how they relate to existing frontier AI safety frameworks (Section 2.3).

2.1 What is a safety case?

A safety case is a structured argument, supported by evidence, that a system is safe enough in a given operational context (MoD, 2007).⁸ Safety cases are often used to inform major "go/no go" decisions, such as whether to build, deploy, procure, or license a system. They typically inform such decisions via a three-staged process, illustrated in Figure 1. First, the safety case is produced by the developer. Then, it is reviewed by an internal or external actor. Finally, it is shared with key decision-makers. Sometimes, the same actor (e.g. a regulator) acts as both reviewer and decision-maker. In some contexts, a safety case is produced already before a system is developed and continually assessed and updated throughout its lifecycle.

Safety cases are a framework for assessing and communicating about the safety of a system. Relative to its alternatives, the defining feature of a safety case is that it is an *argument* about what *outcomes* have been achieved – namely, that the system does not pose unacceptable risk. A safety case is not just a collection of decision-relevant information, nor is it a checklist of best practices followed by the developer. Rather, a safety case explains why the information presented provides sufficient assurance that the system is safe enough (Kelly, 2017).

Safety cases have become increasingly common in recent decades because they are seen as more comprehensive and flexible than traditional approaches to assurance. They emerged in sectors such as energy, petrochemicals, and transportation, where assurance traditionally meant adhering to specific product design rules (Leveson, 2011). In response to accidents in the 1960s-80s, this rules-based approach was questioned on the grounds that it did not encourage developers to comprehensively assess the safety of their systems (Kelly, 2017). Safety cases emerged as an alternative approach. Since then, safety cases have spread to a variety of industries, including defense (David, 2018), aerospace (Dezfuli et al., 2014), and more recently AI-related industries such as security (Alexander et al., 2017), software (Islam & Storer, 2020), and autonomous vehicles (Favaro et al., 2023). However, safety cases have also been criticized for providing a false sense of assurance when in reality it is very difficult to produce an adequate safety case or review it effectively (Langari &

⁸We use the term "operational context" in a broad sense, meaning roughly "setting in which the model is used". In the context of frontier AI, key elements of the operational context include who can access the model and in what ways. Safety cases for frontier AI do not have to be restricted to a specific use case; in fact, the relevant operational context will likely often be a widespread deployment context in which many users can use the system in open-ended ways.

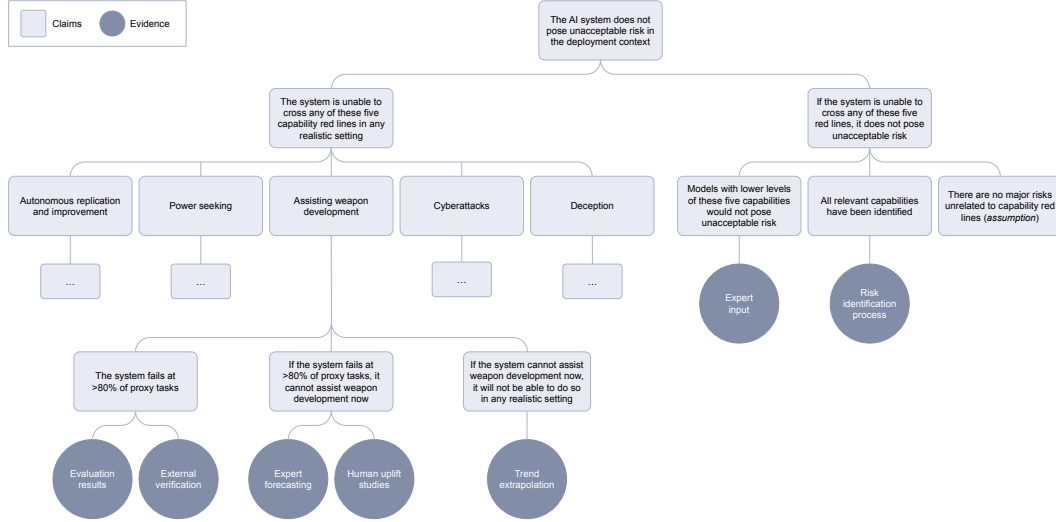


Figure 2: Sketch of a safety case argument

Maibaum, 2013; Leveson, 2011; Wassylng et al., 2011). While there is little empirical evidence for the efficacy of safety cases (Habli et al., 2021), and such evidence is difficult to produce given the nature of risk (Koopman, 2022), practitioners mostly consider them effective (Rinehart et al., 2017), and they are a recognised best practice in the UK (The Health Foundation, 2012).

A safety case has four key components: Objectives, arguments, evidence, and scope. It aims to show that certain safety objectives have been achieved (e.g. that risk is below a specific threshold). The safety case must present one or more arguments for why those objectives have been achieved, with the assumption that the system is unsafe unless convincing arguments are made to the contrary. These arguments must be supported by evidence (e.g. test, evaluation, validation, and verification results). Finally, the safety case must specify the scope in which it is valid (e.g. a specific deployment context with specific safety measures in place). Table 1 outlines these four components and their significance for the unique assurance approach of safety cases.

2.2 What is a frontier AI safety case?

To help give an intuitive sense of safety cases, we sketch what early-stage safety cases for a current frontier AI system may look like in Figure 2 and Table 2. As is typical for current frontier AI systems, the operational context is one of widespread deployment, where many users can interact with the system in an open-ended way. The sketch follows developers’ existing safety frameworks in focusing on catastrophic risks enabled by certain dangerous capabilities. The sketch uses an “inability argument” (Clymer et al., 2024; Goemans et al., forthcoming). It claims that the system does not pose unacceptable risk because it is not capable enough to cause serious harm even if it attempts to. This claim is then broken down into a number of capability red lines (IDAIS, 2024) and providing evidence that the system does not cross these red lines. Note that this sketch is only meant to illustrate what a safety case might look like. We do not claim that a safety case with this structure or substance would be sufficient or sound, even for near-term systems.

2.3 How do safety cases relate to existing frontier AI safety frameworks?

Safety frameworks have been central in the conversation about frontier AI risk management so far. A safety framework is a developer’s plan for assessing and mitigating risks posed by its AI systems (DSIT, 2024). It is an organization-level policy that applies to all frontier systems, in contrast to a safety case which only applies to an individual system. 16 companies have committed to publishing safety frameworks ahead of the 2025 AI Action Summit (DSIT, 2024) and four companies have already done so (Anthropic, 2024; OpenAI, 2023; Google DeepMind, 2024; Magic, 2024). Existing safety frameworks focus on identifying dangerous capabilities, setting capability thresholds above

Components	Explanation	Examples
Scope	<p>System specification: Information about the frontier AI system and deployment set-up (e.g. architecture, training process, safeguards, access).</p> <p>Assumptions: Conditions under which the safety case holds (e.g. temporal scope, assumptions about how the system will be used, assumptions about the effectiveness of safeguards).</p>	<ul style="list-style-type: none"> • Deep learning network with 1 trillion parameters. • Pre-training with 10^{26} total training FLOP using self-supervised learning on CommonCrawl web text. • Fine-tuned with reinforcement learning from human feedback (RLHF) to be helpful, harmless, and honest. • Model will be accessible via an API to anyone with a free online account. • An AI assistant will monitor activity and flag suspected violations of our usage policy to a member of staff. • We include a buffer of expected capability improvements from (a) one year of expected capability improvements from scaffolding and prompting, and (b) additional training using over 10% of pre-training compute. • Once these buffers have been surpassed, a new safety case must be prepared.
Objectives	Safety requirements that the safety case will argue have been met.	<ul style="list-style-type: none"> • The AI system does not pose unacceptable risk in the deployment context. Unacceptable risk is defined as a probability of $\geq 10^{-7}$ / year of causing an event with $\geq 1,000$ fatalities.
Arguments	A hierarchy of claims that, if true, collectively imply that the objectives have been met; an explanation of the developer’s level of confidence that each claim is true and that the objectives have been met.	<ul style="list-style-type: none"> • See Figure 2.
Evidence	Collection of all evidence sources that support the claims made in the argument; detailed explanation of how the evidence was produced, its results, and verification.	<ul style="list-style-type: none"> • Risk identification report covering methodology, risk taxonomy, and external verification. • Results of expert consultation to determine capability red lines covering methodology, names of experts, and results. • Evaluation report covering methodology, results, and the evaluation script. • Report produced by external red-teamers. • Trend extrapolation analysis forecasting effect of post-deployment enhancements on capabilities.

Table 2: Sketch of each safety case component

which additional mitigations would be required, explaining how capabilities will be measured, and outlining mitigation options (Alaga et al., 2024; METR, 2024).

A safety framework could form the basis of a safety case. For example, the central argument of a safety case could be that a system does not cross any of the capability thresholds identified in the developer’s safety framework (similarly to Figure 2), or that the developer has implemented the risk mitigations promised in the safety framework. However, to constitute a safety case, the argument must go beyond simply verifying that the developer has adhered to its framework. The safety case must also justify the framework, explaining why adhering to it implies avoiding unacceptable risk. For example, the safety case must justify the choice of capability thresholds and argue that systems below the thresholds are safe enough to deploy in the chosen way.

Safety cases can play three roles in relation to safety frameworks: (1) Safety cases can be used to document that a developer has adhered to its safety framework with respect to a specific system.

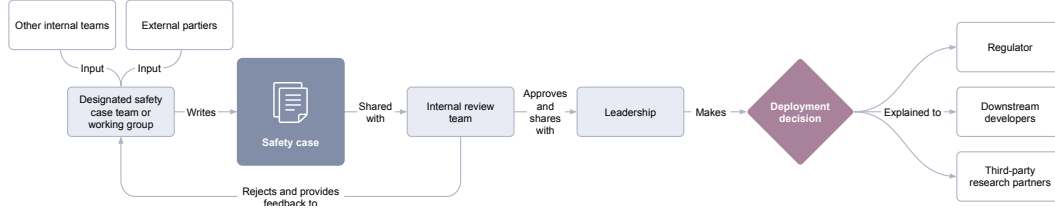


Figure 3: Sketch of an internal safety case review process

Such documentation is an important part of a safety framework. (2) Safety cases can allow safety frameworks to build in more flexibility. They are suitable for supporting subjective outcome-based claims such as “the safeguards are sufficient to prevent unacceptable risk”. They thus allow safety frameworks to use more outcome-based claims in cases where it is difficult to specify adequate pre-commitments. For example, highly specific pre-commitments may be inappropriate for more capable future systems, given that such systems are poorly understood and safeguards are rapidly evolving (3) Safety cases can help make safety frameworks more justified. They require developers to make implicit assumptions in safety frameworks explicit (e.g. that a system below certain capability thresholds does not pose unacceptable risk). While a good safety framework should contain such justification (Alaga et al., 2024), safety cases provide a helpful prompt. Safety cases can thus both draw on prior work on safety frameworks and play a valuable role in their implementation.

3 The case for safety cases

This section outlines the benefits and downsides of using safety cases in frontier AI governance. We consider two contexts in which safety cases could be used: industry self-regulation (Section 3.1) and government regulation (Section 3.2).⁹ We conclude that, in both contexts, safety cases could be a useful and worthwhile tool for the highest-risk systems.

3.1 Safety cases in self-regulation

If developers use safety cases internally as part of self-regulation,¹⁰ they could serve three purposes. First, safety cases could inform major decisions such as starting a training run or deploying a system. Senior management needs to assess if such decisions pose unacceptable risk; safety cases could be a key input into this assessment. They could be produced by a designated safety case team, evaluated and approved by an internal review team, and then shared with leadership who make the decision. A potential internal review process is outlined in Figure 3. Second, safety cases could be used in ongoing risk management. Throughout the lifecycle, the development team can use a safety case as a framework for assessing how safe the system is and what additional safeguards are needed. Third, safety cases can be used to build trust with downstream developers, users, and governments, by providing assurance that the system is safe enough. Many companies in other industries, such as aerospace (Morris & Beling, 2001) and autonomous vehicles (Favaro et al., 2023), use safety cases for all three purposes. They contrast primarily against a less structured approach of sharing individual pieces of evidence or information.

One benefit of using safety cases internally is that they are an effective and scalable way to assess if a system is safe enough. Safety cases provide a way to integrate different sources of evidence into a single, comprehensive assessment of a system’s safety (The Health Foundation, 2012; Rinehart et al., 2017). Moreover, safety cases are flexible enough to be applicable to a wide range of systems. They do not require specific risk assessment or mitigation techniques, but rather invite the developer to use whatever methods are most suitable to a particular system. This is particularly beneficial in the frontier AI context, as future systems may require very different types of assurance arguments to current systems (see Section 4).

⁹Safety cases could also be used in other contexts such as procurement. For example, safety cases are required by the UK Ministry of Defence for all equipment acquisitions (MoD, 2007).

¹⁰By “self-regulation”, we mean companies’ internal policies, voluntary commitments or agreements, and industry-wide guidelines and processes (see (Baldwin et al., 2011; Coglianese & Mendelson, 2010).

Another benefit of using safety cases internally is that they are a useful tool for thinking and communication. Safety cases make otherwise implicit arguments explicit, which has numerous benefits. First, it can highlight reasoning flaws or assurance gaps ([The Health Foundation, 2012](#); [Sujan et al., 2016](#)). This is particularly important in the frontier AI context because risk assessment is still novel and therefore liable to contain errors ([Schuett, 2024](#)). Second, it makes it easier to see how overall system safety is affected if the evidence changes (e.g. if a safeguard stops being effective) ([Inge, 2007](#)). This is particularly important in the frontier AI context because new capabilities, use cases, or risks may emerge after deployment ([O’Brien et al., 2024](#)). Third, it can help stakeholders communicate more clearly about disagreements (e.g. about the validity of an argument or the strength of evidence) by creating a shared framework and set of assumptions ([The Health Foundation, 2012](#); [Rinehart et al., 2017](#)). Fourth, it can aid communication by explaining the context and relevance of the evidence for a system’s safety, facilitating an understanding of both *if* and *when* the system is safe. This can be critical for decision-makers who are not intimately familiar with the system.

One downside of using safety cases internally is that they add costs of two kinds. First, safety cases may add risk management costs, in that they presuppose a reasonably high standard of risk management. However, these costs only apply to developers who do not already meet such a high standard. For example, safety cases should not add significant risk management costs for the companies that already meet their commitment to effectively and transparently manage risks posed by their systems ([DSIT, 2024](#)). Second, safety cases add documentation costs in terms of keeping records of relevant information and structuring this information into an explicit argument. These costs can be reduced with measures such as safety case templates ([Bloomfield et al., 2021](#)) and cross-industry guidance and standards ([UL, 2023](#)).

3.2 Safety cases in regulation

Safety cases could also play several roles in frontier AI regulation¹¹ First, safety cases could support information sharing and transparency (see [Kolt et al., 2024](#)). Regulators could require that developers submit safety cases when deploying frontier models, without regulators having any formal powers to penalize or restrict deployment. This could be considered an extension of existing reporting requirements ([The White House, 2023](#)). Such information sharing could help regulators build capacity for effectively governing AI more broadly. For example, they could help regulators write rules and standards for AI systems behind the frontier by keeping them informed about state-of-the-art safety practices. An information sharing requirement could also be a flexible interim measure, allowing regulators to build experience with safety case review while only scaling up enforcement powers if there is evidence that future frontier models pose serious risk.

Second, safety cases could be used to assess compliance with regulation requiring frontier AI developers to manage risks posed by their systems. Such regulation is already in place in the EU ([European Parliament, 2024](#)) and may also be adopted in other jurisdictions. Regulators could take two broad approaches to assessing compliance with such regulation. First, they could set specific rules about how developers should assess and mitigate risk (e.g. specific safety tests or training techniques that must be used). Alternatively, they could allow developers to choose how to assess and mitigate risk, and use safety cases to assess if developers have done enough.¹² This is a common use of safety cases in industries such as nuclear power ([ONR, 2021](#); [NRC, 2007](#)), on-shore and off-shore petrochemical installations ([UK Parliament, 2015a, 2015b](#)), and rail operators ([Jovicic, 2009](#); [UK Parliament, 2000](#)). In these industries, a regulator must approve a safety case before granting a license to build or operate the relevant facilities. However, safety cases could also be combined with other enforcement mechanisms such as liability. For example, regulators could use safety cases to retrospectively assess if a developer was negligent in deploying a certain model.

Using safety cases to assess compliance with regulation has both advantages and disadvantages relative to using specific rules. These are summarized in Table 3. There is no clear-cut answer to which approach is preferable in the abstract. Rather, this depends on the specific characteristics of the regulator and the industry being regulated. Table 4 outlines some of the relevant characteristics

¹¹By “regulation”, we mean laws, rules, and guidelines set by governments; industry self-regulation is not included. For more information on frontier AI regulation, see [Anderljung et al. \(2023\)](#) and [Schuett et al. \(2024\)](#).

¹²Regulators could also assess compliance by requiring a variety of information from developers without requiring that it be organized as a safety case. The relative benefits and downsides of using safety cases are similar to those outlined in Section 3.1.

		Specific rules	Safety cases
Benefits of safety cases relative to rules	Flexibility and durability	It may be hard for regulators to specify a set of adequate safety practices. Rules might get outdated and be hard to update, locking in suboptimal safety practices.	Regulators can require safety cases without having a clear sense of what specific safety practices are adequate for current or future systems.
	Innovation	Developers lack incentives to develop safety practices beyond what is codified.	Developers are incentivised to develop and use more effective and efficient safety practices as this will allow them to produce better safety cases. However, this requires that regulators are capable of reviewing novel practices.
	Allocation of responsibility	Regulators are responsible for identifying adequate safety practices, which may be difficult given they have more limited information, resources, and AI expertise.	Developers are responsible for specifying adequate safety practices, which is efficient given that they tend to have more information, resources, and AI expertise.
Downsides of safety cases relative to rules	Transparency	Rules are typically public, clear-cut, and manageable in length, making them easier for third parties to scrutinize.	Safety cases may not be fully public, and may be long and technical. This may make it difficult for third parties to scrutinize whether regulators are reviewing safety cases appropriately. However, regulators can facilitate scrutiny by publishing guidance on how they review safety cases, explaining individual decisions, or subjecting themselves to auditing.
	Consistency of enforcement	Compliance is assessed in a consistent, objective way.	Regulator judgment plays a key role in assessing compliance.
	Cost for developer	It is often less costly for developers to demonstrate compliance with rules, and rules provide more legal certainty. However, if rules are poorly specified, they could lead to higher compliance costs by requiring inefficient or unnecessary safety practices.	It is resource-intensive for developers to produce safety cases. Because safety case review is subjective, safety cases also introduce more legal uncertainty for developers.
Factors that could be either a benefit or downside of safety cases relative to rules	Gamability	Rules might create loopholes or gaps.	Because safety case review is subjective, inadequate safety cases might sometimes pass review.
	Risk of regulatory capture	A rulemaking process may be an easier target for industry lobbying relative to a safety case review process.	The subjective judgements required by safety case review might leave more room to give undue favors. Regulators may also rely more on industry for information about safety practices and develop less independent expertise when they do not need to specify adequate safety practices.
	Cost for regulator	It is easier to verify compliance, as it is typically relatively clear-cut if a rule has been followed or not. However, it may be more resource-intensive to develop the rules in the first place and to ensure they are up-to-date.	It is resource-intensive for regulators to verify compliance as it requires a bespoke and subjective review. However, set-up costs might be lower as it might be easier to describe an adequate safety case than to set specific rules.

Table 3: Benefits and downsides of safety cases relative to specific rules

Conditions ¹³	Why are safety cases preferable in this condition?	How likely is it that this condition holds for frontier AI?
Difficult to specify rules: Experts have little confidence in what precise actions will reliably reduce risk to an acceptable level.	Rules are more likely to be lacking and could therefore lock in insufficient or unnecessary safety practices. If there are not well-established best practices, it is also especially important to incentivise innovation in safety practices.	Likely: Frontier AI systems are novel and rapidly evolving; highly general-purpose; complex and not mechanistically understood; heterogeneous and lacking in standard designs. These factors make it hard to specify rules.
Rule lock-in: Once rules have been established, changing them is a difficult and lengthy process.	Rules are more likely to become outdated and lock in suboptimal safety practices.	Unclear: Depends on jurisdiction.
Information asymmetry: Developers have significantly better access to relevant information and expertise compared to regulators.	It is more efficient to place the burden on developers to identify adequate safety practices.	Likely: AI expertise is naturally concentrated in frontier AI companies, though several governments have established AI Safety Institutes to boost their capacity. Developers naturally have more and earlier information about frontier AI, and some tacit knowledge may be difficult to convey even with information sharing requirements.
High risk: The regulated activity poses high risk, defined in terms of the likelihood and impact of the product causing harm.	It is more likely worth imposing a higher compliance cost on developers to reduce risk.	Possible: There is significant expert disagreement about the risks posed by frontier AI. However, it seems difficult to rule out that future AI systems may pose high risk.
Few covered developers: There are few developers and systems that are in scope for the safety case requirement.	The higher compliance cost is limited to a few developers; the enforcement cost for the regulator is lower; and safety case review is likely to be more prompt and thorough.	Likely: Frontier AI development is currently resource-intensive in terms of compute, energy, capital, and specialized expertise. There are therefore only a handful of companies developing frontier AI. However, it is unclear if this will continue to be true.
Covered developers can absorb costs: Covered developers are more well-resourced and able to bear the compliance costs and legal uncertainty of safety cases.	It is more likely worth imposing a higher compliance cost on developers and less likely that compliance costs will stifle innovation.	Likely: Because frontier AI development is resource-intensive, only relatively well-resourced companies are likely to be covered.

Table 4: Conditions under which safety cases are preferable to precise rules

and discusses the extent to which they hold true in the context of frontier AI regulation. Both tables are based on the literature on safety cases ([The Health Foundation, 2012](#); [Hawkins et al., 2013](#); [Inge, 2007](#); [Rinehart et al., 2017](#); [Sujan et al., 2016](#)), and principles-based regulation more broadly ([Coglianese & Starobin, 2020](#); [Decker, 2018](#); [Schuett et al., 2024](#)).

¹³Some conditions that seem relevant, but do not clearly favor either safety cases or specific rules, include: (1) Regulator capacity: Rules seem to require more regulator capacity at the point of writing the regulations, and principles at the point of enforcing the regulations; (2) Degree of alignment between regulator and regulated objectives: If the regulated entities are incentivised to try to circumvent the regulatory objective, it is unclear if rules (which may have gaps and loopholes) or safety cases (which are subjectively assessed) are easier to “game”.

We conclude that the conditions of the frontier AI industry favor the use of safety cases to assess when deployment poses unacceptable risk. First, given that the frontier AI industry is so rapidly evolving, complex, and poorly understood, a more flexible approach to assessing safety seems suitable. Second, safety cases could capitalize on the expertise and resources of frontier AI developers. Third, given that frontier AI systems are currently produced by only a handful of relatively well-resourced developers, the compliance costs seem bearable. However, regulators may want to complement safety cases with specific rules in some cases. For example, regulators could specify rules in better-understood risk domains or mandate some particular safeguards. Regulators could also continuously replace safety cases with precise rules as the frontier moves, using safety cases to identify best practices that are then applied to future systems at a similar capability level. Finally, regulators should consider the implementation challenges outlined in Section 5 when deciding whether to use safety cases.

4 Components of a safety case

This section outlines the content of frontier AI safety cases. We cover the four key components of a safety case: scope (Section 4.1), objectives (Section 4.2), arguments (Section 4.3), and evidence (Section 4.4). For each of these components, we discuss what is feasible today and what developments could happen in the future. We conclude that developers can already produce safety cases with existing techniques, but that research breakthroughs will likely be needed to produce compelling safety cases for more capable future systems.

4.1 Scope

The scope is the range of conditions under which the safety case holds.¹⁴ The scope should include: (1) A detailed specification of the system (e.g. its architecture, training process, and safeguards) and intended deployment context (e.g. whether the model weights will be released or the model will be accessible only via API). (2) An overview of which changes to the system (e.g. post-deployment fine-tuning) or deployment context (e.g. expanding access) are considered within the scope of the safety case and which would require an updated safety case. (3) Any other scope restrictions, in particular the temporal scope (e.g. the safety case being valid for two years), assumptions (e.g. a safeguard being resistant to jailbreaking), or out-of-scope use cases (e.g. use in medical diagnosis).¹⁵ The scope is relevant for reviewers who should assess if the safety case rests on reasonable assumptions. It may also be relevant to audiences such as governments or downstream developers in avoiding unsafe uses or addressing risks not covered by the safety case.

A particular challenge of frontier AI safety cases is that their scope will tend to be broad. Frontier AI systems are highly general-purpose and have so far been deployed widely for open-ended use. This makes it challenging to assure safety, as it is virtually impossible to assess all possible risk scenarios individually (Anwar et al., 2024). As frontier AI systems become more capable, assurance may become even more difficult (see Section 4.3). As such, safety cases may need to be restricted to a narrower scope (e.g. producing cases for specific applications rather than general release). On the other hand, risk assessment and mitigations may advance enough that it is possible to maintain a broad scope. Insofar as assurance is possible, it is both in the commercial interest of developers and the interest of society for developers to widen the scope of safety cases so that systems can be used for a wider range of beneficial applications.

4.2 Objectives

Objectives are requirements that operationalise what it means for the system to be safe enough to deploy (Kelly, 2017). A typical objective is a risk threshold, such as a specified probability of a given level of harm (e.g. a probability of $\geq 10^{-7}$ / year of causing an event with $\geq 1,000$ fatalities). It is common in other industries to set multiple risk thresholds: One region for “unacceptable risk” which must not be exceeded, one region in which risks must be “as low as reasonably practicable”, i.e. where all reasonable mitigations must be implemented, and one region for “acceptable” risk which does not require further mitigations (Koessler et al., 2024). Objectives can also be comparative. For

¹⁴Our definition of “scope” is based on the definition of “context” in Kelly (2017).

¹⁵The safety case may need to address how the developer will prevent the system from being used for out-of-scope purposes.

Box 1: Safety cases throughout the system lifecycle

According to best practice, a safety case is not just a one-off event, but a living document to be developed and maintained across a system's lifecycle (Kelly, 2017; Koopman, 2022). This is illustrated in Figure 4.

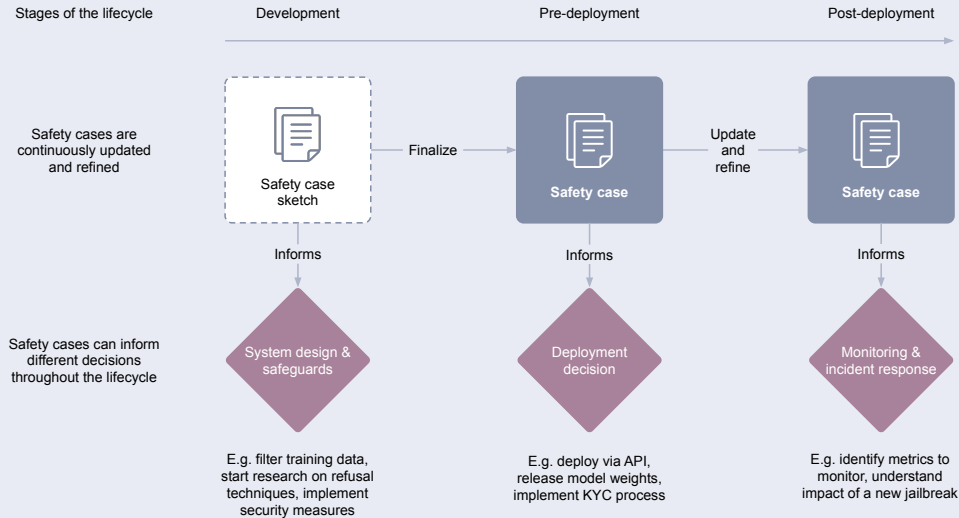


Figure 4: The stage and role of safety cases at each stage of the system lifecycle

During development (e.g. system design, pre-training, and fine-tuning), the developer should gradually build the safety case. The goal is to ensure that safety is embedded in the design of the system rather than tagged on at the end, and to make it easier to produce the pre-deployment safety case. The developer should sketch the intended safety case argument and use it to identify model-level safeguards (e.g. via data filtering or reinforcement learning). The sketch can also help identify potential gaps where more work is needed to assure the safety of the system. Finally, the developer should keep records of evidence that may be used in the safety case. In industries where safety cases inform licensing decisions, such as nuclear energy (ONR, 2021), it is common for the regulator to be involved throughout the process, providing guidance and feedback on early drafts.

After deployment, the developer should continue to update the safety case. New risks may emerge as the system is modified or used in unexpected ways, and real-world information may invalidate the developer's risk assessment. Developers should therefore track key metrics or risk indicators that provide evidence for claims in the safety case (Koopman, 2022). For example, if the safety case estimates that a harm refusal technique has a certain success rate, the developer should track the actual success rate after deployment. Developers should also identify conditions that would require an update to the safety case. Updates could happen at regular intervals and/or in response to triggers such as a new jailbreak or significant post-deployment enhancements.

example, a common objective in autonomous vehicles safety cases is that a self-driving car must be at least as safe as a human driver (Koopman, 2022). In a regulatory context, objectives are typically set by regulators, but can also be selected by developers. In the latter case, reviewers should assess the choices of objectives.

Over time, the objectives of frontier AI safety cases could become more specific, quantitative, and directly based on risks. Early safety cases may initially use broad, qualitative objectives (e.g. that

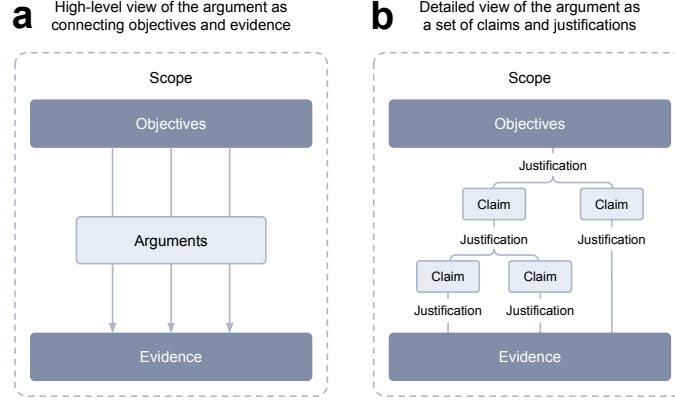


Figure 5: Definition of a safety case argument

the system does not pose unacceptable risk in the deployment context). Objectives could also be restricted to only a few key risk domains, rather than global risk. Finally, early safety cases could use proxy-based objectives, which do not measure risk directly but measure other outcomes indirectly related to risk, such as capability thresholds (Koessler et al., 2024). As safety case methodology develops, objectives should likely move towards direct measurements of risk. They could also increasingly contain sub-objectives for specific risks as risk analysis improves, as is done in nuclear energy (NRC, 1986).

4.3 Arguments

Arguments explain why the evidence gives sufficient reasons to think that the objectives have been met (Kelly, 2017).¹⁶ Arguments break the objective down into subclaims that, if true, collectively imply that the objective has been met, and they explain what evidence supports the subclaims. This is illustrated in Figure 5. For a concrete example, see Figure 2. Reviewers must assess if the argument is valid (i.e. the subclaims imply that the objective has been met) and if the subclaims are adequately supported by the evidence.¹⁷

There can be many ways to argue for the same objective and developers can choose which argumentative strategy to use. For example, safety case arguments could be based on direct risk estimates, comparisons with other systems, or guidelines, as sketched in Figure 6a.¹⁸ Within explicit risk estimates, developers can also adopt multiple strategies. For example, developers can argue that a given system is not capable enough to cause serious harm (inability), that control measures prevent it from causing serious harm (control), that it reliably does not cause serious harm (trustworthiness), or that other AI systems have verified that it will not cause serious harm (deference) (Clymer et al., 2024).¹⁹ Figure 6b contrasts an inability and a control argument. While developers choose

¹⁶The term “argument” is used in a number of different ways in the safety case literature. For example, in the Claims Arguments Evidence framework, it refers to the justification linking a specific claim to a specific piece of evidence (Adelard, 2024); and in propositional logic, an argument would include the evidence and conclusion (i.e the claim that the objective has been met). We follow the use of the term in the Goal Structuring Notation framework (SCSC, 2021).

¹⁷Claims are rarely ever proven or demonstrated with certainty; rather, evidence provides some reason to think the claim is true, but these reasons may be overridden by defeaters (Bloomfield et al., 2021). Some approaches to safety cases model uncertainty more explicitly and systematically (Bloomfield & Bishop, 2010; Denney et al., 2011; Nešić et al., 2021; Wang et al., 2018).

¹⁸These three categories are inspired by the three risk acceptance principles of the European Rail Agency, i.e. the allowed three methods for demonstrating that the risk of a given railway system is acceptable (Jovicic, 2009).

¹⁹Capabilities are far from the only risk factor, and safety cases may also need arguments not closely tied to capabilities, for example about the risk of system failure in high-risk contexts or the diffuse effects of widespread use of the system. However, we focus primarily on capability-related arguments, since those are the focus of developers’ existing safety frameworks (Anthropic, 2024; OpenAI, 2023; Google DeepMind, 2024; Magic, 2024).

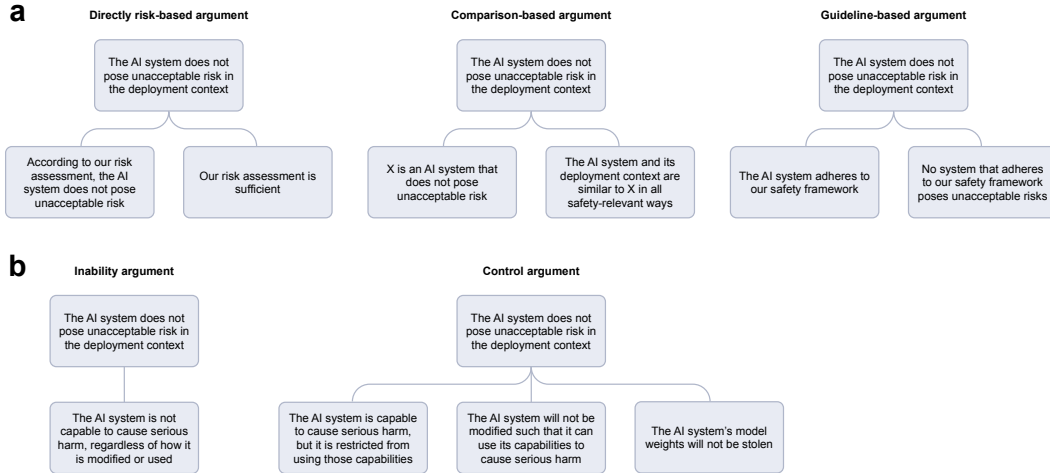


Figure 6: Different types of arguments for frontier AI safety cases

the arguments, regulators may publish guidance on standard arguments they deem acceptable or preferable.

Early frontier AI safety cases will likely rely on inability arguments. An example inability argument, based on existing safety frameworks (as described in Section 2.2), could claim that a system is safe because it does not cross any of a number of capability thresholds. The meat of the argument then consists in justifying the capability thresholds and motivating the internal and external validity of capability evaluations. Inability arguments are relatively well-established. Identifying and evaluating dangerous capabilities is a major focus of existing safety frameworks (Anthropic, 2024; OpenAI, 2023; Google DeepMind, 2024; Magic, 2024), system cards arguably already make inability arguments (Anthropic, 2023; OpenAI, 2024) and researchers are developing early-stage sketches of their explicit structure (Goemans et al., forthcoming). There are still open questions, such as how to incorporate post-deployment enhancements (Davidson et al., 2023), account for defensive uses of capabilities (Mirsky et al., 2023), or address risks that are less closely tied to dangerous capabilities such as systemic risks (Zwetsloot & Dafoe, 2019) or risks from AI malfunction (Raji et al., 2022) that are less closely tied to dangerous capabilities. Nonetheless, it is possible for developers to make reasonable inability arguments already now, and they may be a useful starting point for developing safety case methodology.

In the long run, however, inability arguments are unlikely to be the main focus of safety cases. If frontier AI capabilities continue to accelerate, systems will eventually cross dangerous capability thresholds, and so safety cases must instead argue that those capabilities will not be used to cause unacceptable harm. Other arguments, such as control, trustworthiness, and deference arguments, cannot yet be supported via well-established methods, as described in Section 4.4, and have not yet been stress tested in the real world. By the time safety cases rely on these arguments, the stakes will also be higher, assuming that systems with more dangerous capabilities pose greater risk. For these reasons, developing alternative arguments is a key condition for safety cases to be useful in the long run. Some research on alternative argument structures is already underway (Irving, 2024).

4.4 Evidence

Evidence consists of observable facts that support the claims made in the arguments (Kelly, 2017). Frontier AI safety cases will likely use a wide variety of evidence types. Some examples include *empirical tests* (e.g. dangerous capability evaluations [Phuong et al., 2024; Shevlane et al., 2023], deployment experiments [Weidinger et al., 2023], or simulation exercises to stress test preparedness plans), *mathematical models* (e.g. predictions of post-deployment capability gains [Ganguli et al., 2022]), *formal verifications or proofs* that the system satisfies certain requirements (Dalrymple et al., 2024), *expert judgements* (e.g. results from surveys or interviews), and *documentation of plans, policies, and processes* (e.g. internal governance structures [Schuett, 2023, 2024] or incident

response plans [O'Brien et al., 2024]). Reviewers should both evaluate the relevance and strength of the evidence (i.e. to what extent it supports the claims) and selectively verify individual pieces of evidence (e.g. reproducing select capability evaluations).

There are already a number of techniques available to evidence inability arguments, though some of them are flawed and could be improved with further research. A core challenge of inability arguments is to justify the choice of capabilities and capability thresholds. Current evidence techniques include expert elicitation (Murray et al., forthcoming), threat modeling, public input (The Collective Intelligence Project, 2023), human uplift studies (Weidinger et al., 2023), and forecasting (Phuong et al., 2024). These techniques are relatively early-stage. Over time, safety cases may use more detailed, robust, and externally verified versions of these techniques. For example, theoretical threat models may increasingly be replaced by quantitative estimates backed by empirical evidence. Another core challenge of inability arguments is to assess model capabilities. Developers and third parties have already developed a number of evaluations (Kinniment et al., 2024; Phuong et al., 2024), and system releases tend to be accompanied by internal and external evaluations (Anthropic, 2023; OpenAI, 2024). Again, evaluations are early-stage and there is significant disagreement about their validity (Anwar et al., 2024; Weidinger et al., 2023; Rauh et al., 2024). Nonetheless, it is plausible that these techniques are adequate for current systems, and further iteration could make them more robust.

There is much less scientific understanding of how to evidence arguments beyond inability arguments. It could become challenging to rely on empirical evidence about system behavior as systems may strategically modify their behavior to pass evaluations (Hubinger et al., 2024; Ngo et al., 2022; van der Weij et al., 2024). Safety cases may therefore increasingly need to rely on a mechanistic understanding of systems, for example based on evidence from model internals or the training process (Wasil et al., 2024). Even developers do not currently have such a mechanistic understanding (Anwar et al., 2024). Various research directions may help provide evidence sources for other kinds of safety case arguments. Some examples include mechanistic interpretability (Bereska & Gavves, 2024), formal verification (Dalrymple et al., 2024), control (Greenblatt et al., 2024), and automated AI safety (Bai et al., 2022; Burns et al., 2024; Christiano et al., 2018; Irving et al., 2018; Leike et al., 2018). However, these fields are nascent, it is still unclear if they will produce robust evidence techniques, and doing so will likely require significantly more investment.

5 Implementation challenges

This section discusses potential challenges to using safety cases in frontier AI governance. We discuss two technical challenges, i.e. challenges relating to the content of safety cases (Section 5.1), and three institutional challenges, i.e. challenges relating to the organizations and processes by which safety cases are reviewed (Section 5.2). In each section, we explain how the challenges might affect the use of safety cases in both self-regulation and regulation. Table 5 summarizes the key technical and institutional challenges. If these challenges are left unaddressed, the use of safety cases may be inadvisable. It could result in the conclusions of safety cases being unreliable, biased, and overly relied upon.

5.1 Technical challenges

The first technical challenge is to develop and build consensus on methodology for frontier AI safety cases. This is a particular challenge for three reasons: (1) The idea of applying safety cases to frontier AI is novel and there is not yet any well-established methodology. (2) Frontier AI has many characteristics of complex systems²⁰, which makes assurance particularly challenging. For example, frontier AI is opaque (i.e. not understood at a mechanistic level) and general-purpose (i.e. having a wide and open-ended range of applications). This makes it difficult to apply many known risk assessment techniques (Koessler & Schuett, 2023). (3) There will likely be significant disagreement about what constitutes an adequate safety case. Experts already disagree about how much assurance existing techniques provide. This disagreement will likely grow for future systems. As such, there is a need to invest in methodology for frontier AI safety cases and to work towards consensus on the

²⁰Complex systems are systems that are fundamentally difficult to understand (e.g. due to the interactions of their parts meaning that components cannot be analyzed separately). Complexity has been defined as intellectually unmanagability (Leveson, 2012) or generating outputs with high statistical complexity (Ladyman et al., 2013).

	Technical challenges	Institutional challenges
Self-regulation	<ul style="list-style-type: none"> • Building consensus on safety case methodology • Developing safety cases for more capable future systems 	<ul style="list-style-type: none"> • Implementing an appropriate internal review process • Incorporating third parties into safety case review
Regulation	<ul style="list-style-type: none"> • Setting an appropriate bar for what constitutes an adequate safety case 	<ul style="list-style-type: none"> • Appointing or establishing a body to receive safety cases securely • Building capacity and expertise to effectively review safety cases • Incorporating third parties into safety case review

Table 5: Overview of key technical and institutional challenges to using safety cases to inform decision-making

“bar” a safety case must meet. This is not only necessary for developers in producing high-quality safety cases, but also for reviewers to assess safety cases consistently. Developers are particularly well-placed to address this challenge, but governments can also conduct and coordinate research (Irving, 2024).

However, developers should not delay producing safety cases until this challenge has been resolved. On the contrary, writing safety cases will likely be essential to make progress on methodology. Yet, decision-makers should be aware that early-stage safety cases will likely be somewhat experimental. They should not rely on safety cases until methodologies are more robust. Regulators specifically may also want to delay using safety cases until clearer expectations and standards can be communicated to developers. At the same time, developing standards will be an iterative process and regulators should expect that they will need updating.

The second technical challenge is to develop the safeguards necessary to assure the safety of advanced future systems. At some point, developers may create frontier AI systems with significant dangerous capabilities. As discussed in Section 4, we do not yet know how to reliably assure the safety of such systems. So far, safeguards such as harm refusal have proved easy to circumvent via techniques like jailbreaking (Chao et al., 2023). Safeguards may become even less robust in the future if systems deliberately act to subvert them (Carlsmith, 2023; Hagendorff, 2024; Hubinger et al., 2024; Järvinen & Hubinger, 2024; Pacchiardi et al., 2023; Park et al., 2024; van der Weij et al., 2024). As such, producing safety cases for advanced future systems may require significant breakthroughs in the science of AI safety. Again, developers are likely best-placed to address this challenge, but governments can play a supporting role.

Decision-makers need not wait until this challenge has been resolved before using safety cases to inform decisions. Sufficiently advanced systems should not be deployed until we can be reasonably confident they do not pose unacceptable risk, that is, until it is possible to produce an adequate safety case.²¹ As such, resolving this challenge is not a prerequisite for using safety cases in decision-making, but rather a prerequisite for deploying systems with sufficiently dangerous capabilities.

5.2 Institutional challenges

The first institutional challenge is to implement an appropriate structure for reviewing safety cases. This involves setting a clear process, assigning roles and responsibilities, and creating the right incentive structures for the producers and reviewers of safety cases. This challenge applies both to developers and regulators, though they face different challenges. Developers using safety cases internally should ensure that different organizational functions are responsible for creating, reviewing, and stress-testing safety cases. These responsibilities could be assigned and coordinated using the

²¹In theory, there might be cases in which one can be reasonably confident that a system does not pose unacceptable risk, yet it is not possible to produce an adequate safety case for that system. For example, there may be good reasons to believe that a system is safe enough that cannot be easily written down or communicated. However, it is hard to imagine any such cases in practice.

Three Lines Model, which is a popular risk governance framework that helps organizations to allocate different risk management responsibilities (Schuett, 2023, 2024). Regulators using safety cases for decision-making need to appoint or establish a body to review safety cases. This body must be able to receive and store safety cases securely, given that they will likely contain proprietary and sensitive information. Regulators will also need to consider appropriate checks and balances, such as an appeals process or regular auditing of the review body.

The second institutional challenge is for the review body to build sufficient capacity and expertise to effectively review safety cases. The review body would ideally have expertise in frontier AI systems and risk assessment, capacity to review safety cases in depth, information gathering powers, and model access (Bucknall & Trager, 2023). For developers using safety cases internally, it will be a challenge to balance independence with capacity and expertise. More independent potential reviewers, such as an ethics board (Schuett et al., 2024), may have less time and less hands-on expertise with frontier AI models relative to less independent reviewers, such as an internal safety team. For regulators using safety cases in decision-making, it may be a challenge to attract sufficient expertise. This may require funding to pay higher-than-usual salaries to attract private sector talent (Hopkins, 2012).

While building an effective review system is not a prerequisite for using safety cases in decision-making, it will likely be one of the most important factors in ensuring that they are used well. A core lesson from the other industries is that a capable review body is essential for safety cases to achieve their goals, as safety cases may otherwise be significantly affected by errors and confirmation bias (Haddon-Cave, 2009; Hopkins, 2012; Leveson, 2011; Rinehart et al., 2017). This may be even more true for frontier AI, since the novelty and complexity of the technology means that safety case review may be more subjective than usual. As such, decision-makers should be wary of over-relying on safety cases until the reviewer has built sufficient capacity. However, building capacity will be an ongoing process and practice reviewing safety cases would help develop this capacity. Developers and regulators should therefore consider putting in place provisional bodies that initially primarily provide advice or feedback on safety cases, before eventually adopting decision-making powers. Review guidelines may also help make review more efficient and effective at spotting common pitfalls. Such guidelines are common in other industries that use safety cases (Jovicic, 2009; MAA, 2019; ONR, 2014; UL, 2023).

The third institutional challenge is to include third parties in writing and reviewing safety cases. Including third-party organizations could help address multiple of the challenges raised above, including capacity constraints, gaps in reviewer expertise, and lack of reviewer independence. Third party actors could play several roles. First, they could provide evidence for the safety case itself, such as a governance audit report (Mökander et al., 2023). Second, they could be consulted by the safety case reviewer on specific questions, such as reproducing a model evaluation or reviewing a risk analysis in their domain of expertise. Third, they could review the entire safety case. For example, a developer could commission an independent safety case review to share with relevant decision-makers, such as the board. As another example, a regulator could receive both a safety case from the developer and a “risk case” or red-team of the safety case produced by a third party to help reduce confirmation bias (Clymer et al., 2024).

However, third-party involvement is not a prerequisite for using safety cases in decision-making and need not delay the adoption of safety cases. It is rather a goal to strive for in a mature safety case ecosystem. In addition to determining norms and processes for third party involvement, a key challenge will be to develop the ecosystem in the first place (Birhane et al., 2024; CDEI, 2021). Regulators can support the creation of such an ecosystem (e.g. via funding or other financial incentives).

6 Policy recommendations

Recommendations for developers:

- **Produce safety cases for the next generation of frontier AI systems.** Early safety cases can simply make explicit the arguments that are implicit in current models cards and safety frameworks. As capabilities advance, safety cases may require more rigorous inability arguments (e.g. based on more rigorous risk analysis and externally validated evidence) or other types of arguments (e.g. based on effective safeguards, good governance, or trust in the frontier AI system).
- **Commit to not deploying future generations of AI systems until a safety case has passed internal review.** Developers could also consider making such a commitment already for the next generation of systems given uncertainty about their capabilities and risks.
- **Build capacity to internally produce and review safety cases.** For example, assign roles and responsibilities, hire relevant expertise, set up documentation processes, and determine how safety cases will be reviewed.
- **Start the safety case early in development and continue updating it after deployment.** Developers should sketch the safety case and begin to gather documentation already when planning a new training run. Developers should also continue updating the safety case after deployment, both periodically and in response to specific triggers. These processes can be experimental at first as best practices develop.
- **Share safety cases with external stakeholders, especially governments (e.g. AI Safety Institutes).** Developers should also consider sharing (potentially redacted) versions of safety cases with downstream developers, third party research organizations, and the public.
- **Participate in industry-wide development of safety case methodology and best practices.** For example, participate in industry discussions, share best practices with national AI Safety Institutes, and engage in research partnerships.

Recommendations for governments:

- **Encourage companies to produce and share safety cases.** For example, secure voluntary commitments and set up the necessary infrastructure such as memorandums of understanding with developers and a platform for securely receiving safety cases. In the longer term, consider offering feedback on safety cases.
- **Support companies in implementing safety cases.** For example, convene conversations on safety cases and fund or conduct supporting research such as safety case templates or case studies. By doing so, governments can lower the cost for companies in developing safety cases. In the longer term, engage in a dialogue with industry and third parties to issue guidance on best practices for safety cases.
- **Contribute to the development of a third-party ecosystem to help with producing and reviewing safety cases.** For example, provide funding or other financial benefits to relevant organizations such as auditors or third-party model evaluators.
- **Consider using safety cases to assess compliance with existing or future safety requirements on frontier AI developers.** Take into account industry conditions (Section 3.2) as well as technical and institutional challenges (Section 5). For example, using safety cases is more likely appropriate if capabilities continue to advance rapidly, if basic best practices for safety cases have been established, and if the regulator has the capacity and expertise to effectively review safety cases.

7 Conclusion

This paper has argued that safety cases would be a valuable addition to the toolkit of frontier AI governance, both as a part of industry self-regulation and government regulation. Their first main benefit is that they make an explicit, structured argument, which is helpful for checking if risk assessment is comprehensive and valid. Their second main benefit is they provide a flexible way to assess safety, which is important given that the capabilities, safeguards, and our understanding of frontier AI systems are rapidly developing. It is already feasible to produce rudimentary safety cases based on existing safety frameworks, though significant research breakthroughs will likely be needed to produce safety cases for future systems. While technical and institutional challenges remain, developers and regulators should work to address these and move towards safety case adoption.

Several areas of research could help address the challenges of applying safety cases to frontier AI. First, technical research is needed to sketch safety case arguments and understand gaps where additional research must be conducted before a compelling safety case can be produced. Second, technical and policy research is needed on when and how developers should update safety cases after deployment. There may be lessons from industries such as software and autonomous vehicles. Third, policy research is needed on safety case review processes including the role of third parties, accountability mechanisms, and regulator involvement during development and after deployment.

Safety cases are a promising tool that can be applied already now and scaled to much more capable future systems. However, safety cases for frontier AI face unique challenges. The opaque inner workings and open-ended applications of frontier AI systems make it particularly difficult to predict if and how they will cause harm. Advanced future capabilities could make it extremely challenging to produce adequate safety cases. To solve these challenges, there is an urgent need for companies, governments, and civil society to invest in a collaborative research effort.

Abbreviations

CDEI	UK Centre for Data Ethics and Innovation
COMAH	The Control of Major Accident Hazards Regulations
DSIT	UK Department for Science, Innovation and Technology
ERA	European Union Agency for Railways
IAPS	Institute for AI Policy and Strategy
MAA	Military Aviation Authority
MIT	Massachusetts Institute of Technology
MoD	UK Ministry of Defence
NASA	US National Aeronautics and Space Administration
NCSC	UK National Cyber Security Centre
NRC	US Nuclear Regulatory Commission
ONR	UK Office for Nuclear Regulation
RSCR	The Railways Safety Case Regulations
SCSC	UK Safety-Critical Systems Club
SCSC	Safety-Critical Systems Club
UL	Underwriters Laboratories

Acknowledgments

We are grateful for valuable feedback and comments from Onni Aarne, Ashwin Acharya, Lukas Berglund, Joe Benton, Alexis Carlier, Stephen Clare, Oscar Delaney, Ben Garfinkel, Arthur Goemans, Ben Hilton, Samuel Hilton, Lewis Ho, Gretchen Krueger, Chris Meserole, Joe O'Brien, Anne le Roux, Rohin Shah, Ben Smith, Akash Wasil, Peter Wildeford, Zoe Williams, Peter Wills, and the cohort of the Centre for the Governance of AI's Winter Fellowship 2024. All remaining errors are our own.

References

- Adelard. (2024). *The Adelard safety case development (ASCAD) manual*. Retrieved from <https://perma.cc/27J5-UPQ4>
- Alaga, J., Schuett, J., & Anderljung, M. (2024). A grading rubric for AI safety frameworks. *arXiv preprint arXiv:2409.08751*.
- Alexander, R. D., Hawkins, R. D., & Kelly, T. P. (2017). *From safety cases to security cases*. University of York. Retrieved from <https://perma.cc/434F-NP59>
- Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O’Keefe, C., Whittlestone, J., . . . Wolf, K. (2023). Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.
- Anthropic. (2023). *Model card and evaluations for Claude models*. Retrieved from <https://perma.cc/D8NP-VGL3>
- Anthropic. (2024). *Responsible scaling policy*. Retrieved from <https://perma.cc/DB9F-GAV4>
- Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., . . . Krueger, D. (2024). Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., . . . Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
- Baldwin, R., Cave, M., & Lodge, M. (2011). Self-regulation, meta-regulation, and regulatory networks. In R. Baldwin, M. Cave, & M. Lodge (Eds.), *Understanding regulation: Theory, strategy, and practice*. Oxford University Press. doi: 10.1093/acprof:osobl/9780199576081.003.0008
- Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., . . . Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384(6698), 842–845. doi: 10.1126/science.adn0117
- Bereska, L., & Gavves, E. (2024). Mechanistic interpretability for AI safety: A review. *arXiv preprint arXiv:2404.14082*.
- Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). AI auditing: The broken bus on the road to AI accountability. *arXiv preprint arXiv:2401.14462*.
- Bloomfield, R., & Bishop, P. (2010). Safety and assurance cases: Past, present and possible future – An Adelard perspective. In C. Dale & T. Anderson (Eds.), *Making systems safer* (pp. 51–67). Springer. doi: 10.1007/978-1-84996-086-1_4
- Bloomfield, R., Fletcher, G., Khlaaf, H., Hinde, L., & Ryan, P. (2021). Safety case templates for autonomous systems. *arXiv preprint arXiv:2102.02625*.
- Bucknall, B. S., & Trager, R. F. (2023). *Structured access for third-party research on frontier AI models: Investigating researchers’ model access requirements*. Oxford Martin AI Governance Initiative. Retrieved from <https://perma.cc/E553-EJC2>
- Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., . . . Wu, J. (2024). Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *International Conference on Machine Learning* (pp. 4971–5012). PMLR. Retrieved from <https://perma.cc/5A69-69X8>
- Carlsmith, J. (2023). Scheming AIs: Will AIs fake alignment during training in order to get power? *arXiv preprint arXiv:2311.08379*.
- CDEI. (2021). *The roadmap to an effective AI assurance ecosystem*. Retrieved from <https://perma.cc/CZ9M-HYTC>
- Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krashennnikov, D., . . . Maharaj, T. (2023). Harms from increasingly agentic algorithmic systems. In *ACM Conference on Fairness, Accountability, and Transparency* (pp. 651–666). doi: 10.1145/3593013.3594033
- Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023). Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419v4*.
- Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.
- Clymer, J., Gabrieli, N., Krueger, D., & Larsen, T. (2024). Safety cases: How to justify the safety of advanced AI systems. *arXiv preprint arXiv:2403.10462*.
- Coglianese, C., & Mendelson, E. (2010). Meta-regulation and self-regulation. In R. Baldwin, M. Cave, & M. Lodge (Eds.), *The Oxford handbook of regulation* (pp. 146–168). Oxford University Press. doi: 10.1093/oxfordhb/9780199560219.003.0008

- Coglianesi, C., & Starobin, S. (2020). Management-based regulation. In K. R. Richards & J. van Zeben (Eds.), *Policy instruments in environmental law* (pp. 292–307). Edward Elgar. doi: 10.4337/9781785365683.VIII.20
- Cohen, M. K., Kolt, N., Bengio, Y., Hadfield, G. K., & Russell, S. (2024). Regulating advanced artificial agents. *Science*, 384(6691), 36–38. doi: 10.1126/science.adl0625
- Dalrymple, D. d., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., ... Tenenbaum, J. (2024). Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. *arXiv preprint arXiv:2405.06624*.
- David, R. (2018). *An introduction to system safety management in the MoD (white booklet) Part 1*. MoD. Retrieved from <https://perma.cc/B4LM-F64A>
- Davidson, T., Denain, J.-S., Villalobos, P., & Bas, G. (2023). AI capabilities can be significantly improved without expensive retraining. *arXiv preprint arXiv:2312.07413*.
- Decker, C. (2018). *Goals-based and rules-based approaches to regulation*. UK Department for Business, Energy & Industrial Strategy. Retrieved from <https://perma.cc/YLD3-NFQA>
- Denney, E., Pai, G., & Habli, I. (2011). Towards measurement of confidence in safety cases. In *International symposium on empirical software engineering and measurement* (pp. 380–383). doi: 10.1109/ESEM.2011.53
- Dezfuli, H., Benjamin, A., Everett, C., Feather, M., Rutledge, P., Sen, D., & Youngblood, R. (2014). *NASA system safety handbook. Volume 2: System safety concepts, guidelines, and implementation examples*. NASA. Retrieved from <https://perma.cc/759U-C6E4>
- DSIT. (2024). *Frontier AI safety commitments, AI Seoul Summit 2024*. Retrieved from <https://perma.cc/M9NQ-GNED>
- European Parliament. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Retrieved from <https://perma.cc/J7RE-DRVR>
- Fang, R., Bindu, R., Gupta, A., & Kang, D. (2024a). LLM agents can autonomously exploit one-day vulnerabilities. *arXiv preprint arXiv:2404.08144*.
- Fang, R., Bindu, R., Gupta, A., Zhan, Q., & Kang, D. (2024b). LLM agents can autonomously hack websites. *arXiv preprint arXiv:2402.06664*.
- Favaro, F., Fraade-Blanar, L., Schnelle, S., Victor, T., Peña, M., Engstrom, J., ... Smith, D. (2023). Building a credible case for safety: Waymo’s approach for the determination of absence of unreasonable risk. *arXiv preprint arXiv:2306.01917*.
- Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., ... Clark, J. (2022). Predictability and surprise in large generative models. In *ACM Conference on Fairness, Accountability, and Transparency* (pp. 1747–1764). doi: 10.1145/3531146.3533229
- Goemans, A., Buhl, M., Schuett, J., Korbak, T., Wang, J., Hilton, B., & Irving, G. (forthcoming). *Safety case template for frontier AI: A cyber inability argument*.
- Google DeepMind. (2024). *Frontier Safety Framework Version 1.0*. Retrieved from <https://perma.cc/5Q3D-LM2Q>
- Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843*.
- Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). AI control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*.
- Habli, I., Alexander, R., & Hawkins, R. D. (2021). Safety cases: An impending crisis? In *Safety-Critical Systems Symposium*. University of York. Retrieved from <https://perma.cc/73QJ-XQL2>
- Haddon-Cave, C. (2009). *The Nimrod review: An independent review into the broader issues surrounding the loss of the RAF Nimrod MR2 aircraft XV230 in Afghanistan in 2006*. The Stationery Office. Retrieved from <https://perma.cc/S4HG-9ER9>
- Hagendorff, T. (2024). Deception abilities emerged in large language models. *PNAS*, 121(24). doi: 10.1073/pnas.2317967121
- Hawkins, R., Habli, I., Kelly, T., & McDermid, J. (2013). Assurance cases and prescriptive software safety certification: A comparative study. *Safety Science*, 59, 55–71. doi: 10.1016/j.ssci.2013.04.007
- Helfrich, G. (2024). The harms of terminology: Why we should reject so-called “frontier AI”. *AI and Ethics*, 4(3), 699–705. doi: 10.1007/s43681-024-00438-1
- Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*.
- Hopkins, A. (2012). *Explaining “safety case”*. Australian National University. Retrieved from <https://perma.cc/5VQ8-3FEZ>

- Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... Perez, E. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*.
- IDAIS. (2024). *IDAIS-Beijing, 2024 statement*. Retrieved from <https://perma.cc/RVE3-PLMP>
- Inge, J. R. (2007). The safety case, its development and use in the United Kingdom. In *Equipment safety assurance symposium 2007*. ISSC. Retrieved from <https://perma.cc/T4L4-FP7C>
- Irving, G. (2024). *Safety cases at AISI*. UK AI Safety Institute. Retrieved from <https://perma.cc/6MRR-4VD4>
- Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.
- Islam, G., & Storer, T. (2020). A case study of agile software development for safety-critical systems projects. *Reliability Engineering & System Safety*, 200, 106954. doi: 10.1016/j.res.2020.106954
- Järvinen, O., & Hubinger, E. (2024). Uncovering deceptive tendencies in language models: A simulated company AI assistant. *arXiv preprint arXiv:2405.01576*.
- Jovicic, D. (2009). *Guide for the application of the Commission Regulation on the Adoption of a Common Safety Method on Risk Evaluation and Assessment as Referred to in Article 6(3)(a) of the Railway Safety Directive (ERA/GUI/01-2008/SAF)*. European Railway Agency. Retrieved from <https://perma.cc/C7CC-XXRF>
- Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., ... Narayanan, A. (2024). On the societal impact of open foundation models. *arXiv preprint arXiv:2403.07918*.
- Kelly, T. (2017). Safety cases. In N. Moller, S. O. Hansson, J.-E. Holmberg, & C. Rollenhagen (Eds.), *Handbook of safety principles* (pp. 361–385). Wiley. doi: 10.1002/9781119443070.ch16
- Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., ... Christiano, P. (2024). Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671*.
- Koessler, L., & Schuett, J. (2023). Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. *arXiv preprint arXiv:2307.08823*.
- Koessler, L., Schuett, J., & Anderljung, M. (2024). Risk thresholds for frontier AI. *arXiv preprint arXiv:2406.14713*.
- Kolt, N., Anderljung, M., Barnhart, J., Brass, A., Esvelt, K., Hadfield, G. K., ... Woodside, T. (2024). Responsible reporting for frontier AI development. *arXiv preprint arXiv:2404.02675*.
- Koopman, P. (2022). *How safe is safe enough? Measuring and predicting autonomous vehicle safety*.
- Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, 3(1), 33–67. doi: 10.1007/s13194-012-0056-8
- Langari, Z., & Maibaum, T. (2013). Safety cases: A review of challenges. In *1st international workshop on assurance cases for software-intensive systems (ASSURE)* (pp. 1–6). doi: 10.1109/ASSURE.2013.6614263
- Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. *arXiv preprint arXiv:1811.07871*.
- Leveson, N. G. (2011). *The use of safety cases in certification and regulation*. MIT. Retrieved from <https://perma.cc/LN5W-5UYE>
- Leveson, N. G. (2012). *Engineering a safer world: Systems thinking applied to safety*. MIT Press. doi: 10.7551/mitpress/8179.001.0001
- MAA. (2019). *Manual of air system safety cases (MASSC)*. Retrieved from <https://perma.cc/P3FM-LYNA>
- Magic. (2024). *AGI readiness policy*. Retrieved from <https://perma.cc/AQ5N-G5V5>
- METR. (2024). *Common elements of frontier AI safety policies*. Retrieved from <https://perma.cc/5NN9-QPLU>
- Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., ... Biggio, B. (2023). The threat of offensive AI to organizations. *Computers & Security*, 124, 103006. doi: 10.1016/j.cose.2022.103006
- MoD. (2007). *Safety management requirements for defence systems (Standard 00-56)*. Retrieved from <https://perma.cc/YUP4-V83D>
- Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. *AI and Ethics*. doi: 10.1007/s43681-023-00289-2
- Morris, A., & Beling, P. (2001). Space shuttle RTOS bayesian network. In *20th Digital Avionics Systems Conference*. doi: 10.1109/DASC.2001.963378

- Mouton, C. A., Lucas, C., & Guest, E. (2023). *The operational risks of AI in large-scale biological attacks: A red-team approach*. RAND. doi: 10.7249/RRA2977-1
- Murray, M., Dreksler, N., Schuett, J., Anderljung, M., Rand, A., Marcoci, A., & Garfinkel, B. (forthcoming). *Estimating the marginal risk of LLMs: A pilot study*.
- NCSC. (2024). *The near-term impact of AI on the cyber threat*. Retrieved from <https://perma.cc/6NPT-UHX4>
- Nešić, D., Nyberg, M., & Gallina, B. (2021). A probabilistic model of belief in safety cases. *Safety Science*, 138, 105187. doi: 10.1016/j.ssci.2021.105187
- Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.
- NRC. (1986). *Safety goals for the operations of nuclear power plants: Policy statement republication*. Retrieved from <https://perma.cc/3M2V-U4TK>
- NRC. (2007). *10 CFR § 52.79 – Contents of applications; technical information in final safety analysis report*. Retrieved from <https://perma.cc/P2VS-8UBN>
- O’Brien, J., Ee, S., & Williams, Z. (2024). Deployment corrections: An incident response framework for frontier AI models. *arXiv preprint arXiv:2310.00328*.
- ONR. (2014). *Safety assessment principles for nuclear power plants*. Retrieved from <https://perma.cc/5RQK-YCD8>
- ONR. (2021). *Licensing nuclear installations*. Retrieved from <https://perma.cc/B762-8SYY>
- OpenAI. (2023). *Preparedness Framework (Beta)*. Retrieved from <https://perma.cc/9FBB-URXF>
- OpenAI. (2024). *OpenAI o1 system card*. Retrieved from <https://perma.cc/8PA5-UBL5>
- Pacchiardi, L., Chan, A. J., Mindermann, S., Moscovitz, I., Pan, A. Y., Gal, Y., ... Brauner, J. (2023). How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. *arXiv preprint arXiv:2309.15840*.
- Park, P. S., Goldstein, S., O’Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5). doi: 10.1016/j.patter.2024.100988
- Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., ... Shevlane, T. (2024). Evaluating frontier models for dangerous capabilities. *arXiv preprint arXiv:2403.13793*.
- Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). The fallacy of AI functionality. In *ACM Conference on Fairness, Accountability, and Transparency* (pp. 959–972). doi: 10.1145/3531146.3533158
- Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Comanescu, R., Akbulut, C., ... Weidinger, L. (2024). Gaps in the safety evaluation of generative AI. In *AAAI/ACM Conference on AI, Ethics, and Society*. AAAI.
- Rinehart, D. J., Knight, J. C., & Rowanhill, J. (2017). *Understanding what it means for assurance cases to “work”*. NASA. Retrieved from <https://perma.cc/TA87-8TZL>
- Sandbrink, J. B. (2023). Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*.
- Scheurer, J., Balesni, M., & Hobbhahn, M. (2024). Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*.
- Schuett, J. (2023). Three lines of defense against risks from AI. *AI & Society*. doi: 10.1007/s00146-023-01811-0
- Schuett, J. (2024). Frontier AI developers need an internal audit function. *Risk Analysis*, 1–21. doi: 10.1111/risa.17665
- Schuett, J., Anderljung, M., Carlier, A., Koessler, L., & Garfinkel, B. (2024). From principles to rules: A regulatory approach for frontier AI. *arXiv preprint arXiv:2407.07300*.
- SCSC. (2021). *Goal structuring notation community standard (Version 3)*. Retrieved from <https://perma.cc/CD9W-YX6S>
- Seeger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., ... Gupta, A. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*.
- Shevlane, T. (2022). Structured access: An emerging paradigm for safe AI deployment. In J. B. Bullock et al. (Eds.), *The oxford handbook of AI governance*. Oxford University Press. doi: 10.1093/oxfordhb/9780197579329.013.39
- Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., ... Dafoe, A. (2023). Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.
- Soice, E. H., Rocha, R., Cordova, K., Specter, M., & Esvelt, K. M. (2023). Can large language models democratize access to dual-use biotechnology? *arXiv preprint arXiv:2306.03809*.

- Solaiman, I. (2023). The gradient of generative AI release: Methods and considerations. In *ACM Conference on Fairness, Accountability, and Transparency* (pp. 111–122). doi: 10.1145/3593013.3593981
- Sujan, M. A., Habli, I., Kelly, T. P., Pozzi, S., & Johnson, C. W. (2016). Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices. *Safety Science*, 84, 181–189. doi: 10.1016/j.ssci.2015.12.021
- The Collective Intelligence Project. (2023). *Participatory AI risk prioritization*. Retrieved from <https://perma.cc/8P29-NUCH>
- The Health Foundation. (2012). *Evidence: Using safety cases in industry and healthcare*. Retrieved from <https://perma.cc/X38X-EDXP>
- The White House. (2023). *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Executive Order 14110)*. Retrieved from <https://perma.cc/99VP-6H55>
- UK Parliament. (2000). *The railways (safety case) regulations 2000*. Retrieved from <https://perma.cc/849E-3G9Y>
- UK Parliament. (2015a). *The control of major accident hazards regulations 2015*. Retrieved from <https://perma.cc/8B34-36Z3>
- UK Parliament. (2015b). *The offshore installations (offshore safety directive) (safety case etc.) regulations 2015*. Retrieved from <https://perma.cc/H94N-CDVF>
- UL. (2023). *Evaluation of Autonomous Products (UL 4600)*.
- Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3), 189–191. doi: 10.1038/s42256-022-00465-9
- van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). AI sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*.
- Wang, R., Guiochet, J., Motet, G., & Schön, W. (2018). Modelling confidence in railway safety case. *Safety Science*, 110, 286–299. doi: 10.1016/j.ssci.2017.11.012
- Wasil, A. R., Clymer, J., Krueger, D., Dardaman, E., Campos, S., & Murphy, E. R. (2024). Affirmative safety: An approach to risk management for high-risk AI. *arXiv preprint arXiv:2406.15371*.
- Wassylng, A., Maibaum, T., Lawford, M., & Bherer, H. (2011). Software certification: Is there a case against safety cases? In R. Calinescu & E. Jackson (Eds.), *Foundations of computer software. modeling, development, and verification of adaptive systems* (pp. 206–227). Springer. doi: 10.1007/978-3-642-21292-5_12
- Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., ... Isaac, W. (2023). Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv:2310.11986*.
- Yohsua, B., Daniel, P., Tamay, B., Rishi, B., Stephen, C., Yejin, C., ... Tramèr, F. (2024). *International scientific report on the safety of advanced AI: Interim report*. DSIT. Retrieved from <https://perma.cc/PM8K-9G48>
- Zwetsloot, R., & Dafoe, A. (2019). *Thinking about risks from AI: Accidents, misuse and structure*. Lawfare. Retrieved from <https://perma.cc/57GZ-QXBD>