Proceedings Paper:
FEAKINS, SHAUN, HABLI, IBRAHIM and MORGAN, PHILLIP DAVID JAMES (2026) Clear, Compelling Arguments: Rethinking the Foundations of Frontier AI Safety Cases. In: The International Association for Safe & Ethical AI Conference (IASEAI'26).

Clear, Compelling Arguments: Rethinking the Foundations of Frontier AI Safety Cases

Shaun Feakins
UKRI Centre for Doctoral Training in Safe AI Systems (SAINTS)
Institute for Safe Autonomy
Deramore Ln, York YO10 5GH
shaun.feakins@york.ac.uk

Ibrahim Habli
UKRI Centre for Doctoral Training in Safe AI Systems (SAINTS)
Institute for Safe Autonomy
Deramore Ln, York YO10 5GH
ibrahim.habli@york.ac.uk

Phillip Morgan
UKRI Centre for Doctoral Training in Safe AI Systems (SAINTS)
The York Law School
Freboys Ln, York YO10 5GD

Abstract

This paper contributes to the nascent debate around safety cases for frontier AI systems. Safety cases are structured, defensible arguments that a system is acceptably safe to deploy in a given context. Historically, they have been used in safety-critical industries, such as aerospace, nuclear or automotive. As a result, safety cases for frontier AI have risen in prominence, both in the safety policies of leading frontier developers and in international research agendas proposed by leaders in generative AI, such as the Singapore Consensus on Global AI Safety Research Priorities and the International AI Safety Report. This paper appraises this work. We note that research conducted within the alignment community which draws explicitly on lessons from the assurance community has significant limitations. We therefore aim to rethink existing approaches to alignment safety cases. We offer lessons from existing methodologies within safety assurance and outline the limitations involved in the alignment community's current approach. Building on this foundation, we present a case study for a safety case focused on Deceptive Alignment and CBRN capabilities, drawing on existing, theoretical safety case "sketches" created by the alignment safety case community. Overall, we contribute holistic insights from the field of safety assurance via rigorous theory and methodologies that have been applied in safety-critical contexts. We do so in order to create a better foundational framework for robust, defensible and useful safety case methodologies which can help to assure the safety of frontier AI systems.

1 Introduction

Safety cases are used to make clear, defensible arguments that a system is acceptably safe in a given context. This paper considers the nascent approach to safety cases by those involved in and concerned about frontier AI systems, which we term "alignment safety cases". We argue that alignment safety

Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI'26).

cases diverge significantly from many foundational methods found within safety-critical systems methodologies, limiting their effectiveness. We present a critical appraisal of the existing approach to alignment safety cases. We aim to recentre the debate using foundational and technology-agnostic methodologies from the field of safety assurance.

Safety cases have traditionally been concerned with assuring safety-critical systems. Safety-critical systems are those systems whose failure could result in loss of life, significant property damage or damage to the environment [54]. Safety assurance focuses on how to communicate, assess and establish confidence in sufficient risk reduction in high-criticality domains, such as aerospace, automotive, nuclear and chemical contexts [41]. The field, also known as safety science or systems safety, incorporates research areas such as engineering [16], human factors within teams [75], organisational culture [71], and safety argumentation [36]. Safety cases have been used across safety-critical industries for decades. They are best understood within safety science as structured frameworks for thinking about the safety of a system, which should result in a compelling and defensible argument about the system from development through to post-deployment.

Alignment safety cases are a growing field of research which aim to use safety cases to help to assure the safety and alignment of frontier AI systems. Frontier AI systems are defined as highly capable general purpose models which *"match or exceed the capabilities present in today's most advanced models"* [28]. Alignment safety cases often include a particular focus on catastrophic risks [14, 20, 35, 57, 8] (Appendix B). We define the alignment safety case literature as the current research direction and industry-adopted approach to safety cases by frontier AI developers. Clymer et al. and Buhl et al. are examples of initial research in this area, with later work at the U.K. AI Security Institute building directly on Clymer et al.'s framework [35, 57].

This paper begins by presenting an overview of safety cases within safety-critical systems. It moves to illustrate core issues with the alignment safety case literature's understanding of and rationale behind safety cases. We then introduce foundational methodologies within safety assurance, and how these might be applied by the alignment safety case community. Finally, we close with a case study illustrating a basic safety argument about two hazardous events - CBRN capabilities and Deceptive Alignment - presented by frontier AI systems, and how a safety case and associated risk assessment might respond to those hazardous events.

2 Safety-Critical Systems, Safety Cases and Frontier AI Assurance

An increasing body of work advocates for applying methodologies derived from safety-critical industries and safety assurance to frontier AI [81, 22, 11]. On Knight's definition, that safety-critical systems are those systems wherein failure could result in loss of life or significant property damage [54], frontier AI systems fit the basic definition of a safety-critical system in some deployment contexts.

The safety case is a seminal technique within safety-critical systems. Safety cases are accompanied by a rich body of academic and industrial literature and research. A safety case aims to offer *"a structured argument, supported by a body of evidence that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given operating environment."* [62]

Within safety-critical systems, there are various characterising features of safety cases, involving argument structure [36], confidence assessments [37] and a through-life process [62, 41]. Safety case arguments are often presented in graphical notations, such as Goal Structuring Notation (GSN) [52]. Importantly**,** the safety case is a through-life document, developed from the start of system development through to decommissioning [62]. Safety cases can therefore occasionally become lengthy documents, involving both system level and component level analyses [41].

The safety case was first introduced formally in the U.K. for the nuclear industry in 1965 [42]. An increasing number of industries which aim to assure the safety of their systems have adopted safety cases, from ML-based safety-critical systems within automotive [69], through to healthcare [32], high-rise building safety [46], and AI systems themselves [41].

2.1 Alignment Safety Cases

This background to safety cases motivated those involved in assuring the safety and alignment of frontier AI systems [20, 14, 48]. The fact that safety cases have been used in safety-critical settings explicitly form the justification of Buhl et al. and Clymer et al.'s use of safety cases:

Clymer et al.: *"We first introduce the concept of a "safety case," which is a method of presenting safety evidence used in six industries in the UK (Sujan et al., 2016). A safety case is a structured rationale that a system is unlikely to cause significant harm if it is deployed to a particular setting."*

Buhl et al.: *"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)...However, there is still little clarity on what frontier AI safety cases would look like and how they could be used."*

Since 2024, safety cases for frontier AI have received significant interest. Several significant reports, such as the Singapore Consensus on Global AI Safety Research Priorities and the International AI Safety Report [11, 10], use both Clymer et al. and Buhl et al. to define safety cases. This work has also been referred to in both Anthropic and Google DeepMind's inclusion of safety cases in their safety frameworks [4, 24]. Furthermore, the U.K. AI Security Institute (AISI) made safety cases a specific research agenda in 2024 [50]. Researchers at AISI have since built specific safety case "sketches" and "templates" which have built directly on Clymer et al.'s argument subcomponents, such as a "cyber inability argument" [35] and a "sketch of an AI control safety case" [57].

Alignment safety case approaches vary significantly from long-established assurance practices in safety-critical industries, despite borrowing the term, the basic idea behind the methodology and the rationale for safety cases. While it is a novel field, alignment safety case arguments have been presented in order to create justifications post development, upon deployment, as to why systems do not display dangerous capabilities [20] , alongside continuous monitoring post-deployment. Post-development justification of why a system is safe is only part of a broader aim of a safety case within the assurance community, and we illustrate the limitations of this approach throughout this paper.

The alignment safety case work rarely or insufficiently considers that the safety case is a tool for through-life consideration of how to mitigate eventual harms, hazards and risks. Although the research is novel, illustrated by Hilton et al.'s open problems [48], we argue that alignment safety case research is proceeding along a research direction which diverges significantly from the safety-critical techniques it aims to borrow.

2.2 Rationale

There are multiple significant differences between the alignment safety case approach and the safety assurance approach. One might argue that this arises from the novelty of advanced AI systems. However, this paper suggests that lessons from the safety assurance literature remain appropriate to mitigating issues with existing alignment safety cases.

Importantly, if the fundamental methods on which alignment safety cases are built are deeply divergent from those of the assurance community, then the rationale underpinning Clymer et al. and Buhl et al.'s papers for using safety cases breaks down, as both rely on the fact safety cases have been used in safety-critical industries successfully:

Fact: Safety cases have been used in safety-critical settings.

Premise: Safety-critical settings provide useful scenarios and techniques which can be translated to frontier AI systems.

Conclusion: Safety cases should be used to assure frontier AI systems.

Condition: If no relevant safety case techniques from safety-critical settings are translated to frontier AI systems, then the premise underpinning the conclusion becomes irrelevant.

The alignment safety case literature clearly intends to incorporate the condition set out in the above reasoning. Every article surveyed in Appendix B draws on safety-critical methodologies or literature in some form: foundational papers rely on established safety case authors within their arguments

[48, 14, 20]. Similarly, the alignment safety case literature generally aims to use safety case notations and references [8, 35, 21, 13, 57, 56].

We agree with the literature's aims to incorporate novel techniques to the assurance of frontier AI systems. Nonetheless, the motivation of this paper is to introduce relevant concepts that we believe are directly translatable to the nascent alignment safety case literature. Without this grounding, the underlying motivation for using safety cases becomes irrelevant and the concept of presenting evidence upon deployment for the safety of a system should be rephrased.

2.3 Foundational Differences between Alignment and Assurance School

2.3.1 The Use of Assurance Literature

Some work within the alignment safety case community uses safety case literature in ways which might be challenged by those in safety assurance. For example, one paper suggests 'concretising' safety case arguments into hard standards [20]. Hard standards are rigid and static. Safety cases are dynamic documents [26]. Safety cases are not templates which are thereafter hardened into rules. This proposal directly limits the underlying justification of safety cases, which is based in a goal-setting regulatory framework which prioritises flexibility and context-specific safety arguments rather than prescriptive standards [42].

There also appear to be issues within the construction of safety argument notations, a fundamental element of safety case construction [36]. For example, in one Claims, Arguments, Evidence framework, a claim (C2.2: This AI System poses no risk of novel Cyberattack) is supported by evidence instead of arguments [35]. This misses a core aspect of safety case notation, which involves evidence-based explanation and justifications of claims via argumentation [52]. Other approaches base safety case methodologies on ideas which would surprise safety engineers, such as the idea of a 'national security safety case' [68]. Historically, safety cases and security cases have been separate domains (despite efforts to integrate them), particularly within the defence context of national security. This is encapsulated by Alexander et al., which examines the application of assurance cases to the security domain, noting in separate sections that "there are practical challenges in moving to security cases" and that "there are significant differences between safety and [traditional] security" [3].

One influential paper briefly suggests reviewing risk cases alongside safety cases [20], a recommendation which has since been directly cited by the International AI Safety Report [11]. However, risk cases are a method which have been reviewed and rejected by industry authorities, such as the MOD, within safety assurance [32, 70], and were covered in limited detail within the paper [20]. Furthermore, risk cases were initially designed to replace, not complement, safety cases [42, 70].

These issues demonstrate the importance of stronger collaboration between the safety assurance community and those with expertise and deep familiarity with frontier AI safety.

2.3.2 Deployment vs development settings

Many alignment safety cases either refer only to deployment settings or cover development settings in limited detail [21, 35, 20, 48]. This work has to date omitted substantial discussion of foundational through-life elements within systems safety. This might include altering development conditions in response to certain concerns, such as altering pre-training techniques or post-training methods to limit CBRN capabilities [19]. It might involve management considering more abstract potential harms of developing novel architectures prior to development, such as the risks involved in more performant [43] but obfuscated chain-of-thought reasoning [55]. It might also involve the pre-deployment testing outlined by Clymer et al. and others in the literature, as well as steps taken post-deployment through to decommissioning. While alluded to in the literature [35, 8, 48], the focus on through-life evidence has been more limited.

The alignment safety case literature could be viewed as basing its interpretation of safety cases on the idea that a system 'is safe to deploy'. However, the justification for the system being safe to deploy in this literature is not currently due to through-life assurance, as underpins the safety assurance literature (e.g. 'safe' design and 'safe' deployment) [62, 32, 41]. Instead, it is presented as safe because the developer cannot find issues with the system upon deployment. Clymer et al. discuss the safety case as applying to a specific 'deployment window'. They silo this approach by noting that they 'specifically focus on deployment decisions'. While noting that their framework *"could* *also be adapted to decisions about whether to continue AI training"* [20], this necessarily separates interrelated areas of a model's lifecycle [20]. We believe this justification is insufficient: it would be insufficient in any other safety-critical industry for a model developer to test a model on deployment, without consideration of the broader system's earlier decisions and processes, and then claim that it is safe to deploy.

The use of post-development justifications for deployment within the alignment safety case literature indicates a significant divergence from the safety assurance approach. By focussing on deployment decisions, the alignment safety case literature necessarily argues why a system is safe to be released *because it does not do bad things upon deployment*, rather than the fact that a *system is safe because developers have made careful decisions before and throughout the development and deployment lifecycle*. An equivalent distinction may be made between aerospace engines and frontier AI. Within aerospace, systems are safe to deploy because certain engineering decisions have been made throughout the development lifecycle to reduce the chance of physical harm occurring, not because when planes are tested they do not crash. While testing is a part of pre-release, for example by testing the impact of birdstrikes on propellers [30], it forms one of many steps within the safety argument lifecycle.

More recent work, such as Hilton et al., includes discussion of this issue, but offers minimal detail. Buhl et al. briefly discuss through-life steps, drawing on best practice from the safety assurance community. Hilton et al. do mention actionable development steps and engage with the risk assessment process underpinning a safety case, mentioning development and pre-deployment steps within this process. However, this content is left relatively unexplored. For example, Hilton et al. imply that organisational and training culture arguments are only relevant for 'low-capability systems' [48]. In contrast, one might argue that the higher the capability of a system, the more likely it is that the system needs to be deployed within safe organisational contexts in order to mitigate catastrophic risks from opaque systems.

Alternatively, Hilton et al. ask at the end of their paper *"To what extent do current training pipelines incentivize deceptive behaviour?"*. This type of question is pertinent to a safety case, which often deals with uncertainty [1]. However, Hilton et al. split safety cases into separate components, such as a 'pretraining safety case' [48], which would necessarily split the holistic pipeline of safety evaluation into separate components, rather than acknowledging the complexity of frontier systems and the interrelatedness of development steps. This undermines the strength of a safety case in allowing for a holistic evaluation of a system, rather than individual components of a technology.

The omission in alignment safety case literature of through-life considerations within safety assurance may have had an impact on how safety cases are viewed by those in the wider alignment community. For example, Bowen et al., who reference alignment safety case literature and the U.K. Ministry of Defence, effectively strawman the safety case by stating that under the 1997 U.K. Ministry of Defence Standards, *"safety evaluations only need to confirm a single compelling safety case"* [12]. This statement could be a response to either alignment safety cases or the idea of a safety case within safety assurance. Bowen et al. rely on the statement to justify that safety cases are insufficient, as developers only need to present an argument about the model at pre-mitigation or post-mitigation stage, and thereafter *"testing can stop"* [12]. This appears to be a questionable interpretation of Ministry of Defence safety assurance requirements. For example, the Ministry of Defence has various comprehensive documents on through-life capability management [7]. Similarly, the safety evaluation of any safety-critical system does not stop upon release. This is among the most basic principles of safety engineering, supported by a rich body of debate and evidence (e.g. criticised by Kelly in 2008 as '*Safety case shelf-ware*' [53]). When military software or hardware is released, post-deployment strategies for potential issues are constantly reevaluated until decommissioning [7]. It is feasible that Bowen et al. have interpreted safety cases as they do due to the alignment safety cases conducted within the broader field of alignment research. In either case, this issue illustrates the downstream impacts of a misunderstanding of the foundational aims of safety cases.

We aim to respond to and build on these issues and interpretations of safety assurance perspectives to recentre the debate around safety cases. We offer perspectives from safety assurance to illustrate a deeper outline of a risk assessment underpinning a frontier AI safety case, where there may be overlaps between safety assurance risk assessments and frontier AI risk assessments, and how that might inform a safety argument.

3 Risk assessment

Risk management informs and underpins a safety case within safety science. The safety case therefore reflects associated systematic risk management steps. This principle moves safety cases from "Paper Safety" rubber-stamp documents through to a document which helps to assure that systems are safe in a given context. Paper safety represents the idea that when safety cases are done incorrectly, they can simply be confirmatory, bureaucratic exercises which rubber-stamp safety arguments [42]. Existing criticism of safety cases within the alignment community indicates well that nascent work on alignment safety cases may not sufficiently illustrate the risk management process throughout the development lifecycle [38]. This approach could inadvertently fall into the trap of 'paper safety'.

This argument is implicit in Greenblatt's criticism of safety cases for frontier AI, given he suggests that the approach has bad 'epistemic effects' [38]. Greenblatt states that safety cases are not useful tools because developers should instead focus on collecting evidence [38]. In contrast, the focus of a safety case within the safety assurance community should be to collate risk-based evidence throughout the development process. Greenblatt, whether intentionally or inadvertently given his work within the alignment community, echoes our criticism of alignment safety cases by drawing on well-established debates within safety assurance. Furthermore, Bowen et al. implicitly make this criticism of alignment safety cases by interpreting a safety case as a justification that 'confirms' a single document.

The alignment safety case literature evidently does not aim to produce confirmatory documents, given there is significant rigour in the evaluative methods [20, 35, 8]. Instead, we suggest that the idea behind producing a safety case for frontier AI systems, as well as the testing methodologies presented by the wider literature [8, 20, 13], could form important parts of safety evaluation of frontier AI more generally. We aim to ameliorate those issues here.

3.1 Risk and Hazard Identification

The typical way in which risk management is constructed within safety engineering involves identifying hazards and hazardous events, assessing, controlling and monitoring the risk of these hazards, setting out an acceptable level of risk, documenting those risks formally, and placing this information into a safety argument which contextualises and justifies the risk management process and outputs. It is an iterative and through-life process. Safety management within safety assurance always begins with scoping the system and its context and identifying *hazards* or *hazardous events* and *risks* [66].

- *Hazardous event:* An event that can cause harm
- *Hazard*: Potential source of harm
- *Risk:* The combination of the probability of occurrence of harm and the severity of that harm

3.1.1 Hazards

There are various approaches to hazard analysis. Hazards must be approached consistently and defined clearly, which is a key consideration for frontier AI. The identification of hazards or hazardous events underpins safety assurance. There are new kinds of hazardous events presented by increasingly advanced AI. The focus in the alignment safety case literature is on catastrophically dangerous misalignment [20, 8], as well as associated hazards, such as those posed by CBRN capabilities [14]. Both approaches focus squarely on the capabilities presented by frontier AI systems, hence the focus on novel hazardous capabilities. Hazardous events in ISO/IEC Guide 51:2014 appear to be analogous to the term "threat model" used in the alignment literature for risks posed by frontier systems themselves [20].

Nonetheless, model developers will have to identify which hazards or hazardous events are of concern. It is unclear currently which hazardous events should be identified in a frontier AI system [48]. Within safety assurance, hazards are those which could cause physical, environmental or increasingly some psychological harm [31]. Moreover, safety assurance has grappled with the expanding scope of harms in safety-critical systems in recent years, towards sociotechnical understandings of how systems are developed [29]. This shift has been marked explicitly by the introduction of AI-based systems into safety-critical settings [15]. It has therefore been accompanied by considerable debate across the field, building on an extensive body of academic, policy and industry best practice. The question of who decides on which type of hazards should be included is a pertinent question for any suggestion that a regulator will evaluate safety cases [23].

The question of hazards and harms is perhaps more complex than it may initially seem. For example, Clymer et al. suggest that any catastrophic harm is relevant. Other research is motivated by catastrophe caused by scheming or potentially superintelligent models [56]. Clymer et al. define catastrophic harm as 'large-scale devastation of a specific severity'. Their definition involves "billions of dollars in damages or thousands of deaths" [20]. Buhl et al. consider a broader scope of hazards, such as ability for the model to assist with weapon creation [14].

Misalignment could ostensibly cause many harms outside of catastrophic or existential harm. For example, sycophancy could amplify bias or cause certain psychological harms [78]. However, some definitions omit these harms in order to focus their safety analysis onto specific hazards [20]. Similarly, frontier developers have explicitly noted that their duties are not to push the frontier of capabilities without suitable risk management processes [4, 24, 64].

However, there are potentially some inconsistencies in the current approach to hazard analysis. For example, it is well-established that goal misspecification is a hazard [61]. If goal misspecification were present in some cases but not others, does it only become a hazardous event once it reaches the threshold of causing billions of dollars of damages? Or when a certain amount of deaths could be reasonably foreseen? Or is it only a hazard where the system is also superintelligent and is intentionally scheming? If so, who decides? This is where risk management techniques and associated goal-based policies begin to play an important aspect of safety argumentation.

3.1.2 Risk

Within safety management, the relevant and affected stakeholders, e.g. regulators, users and developers, need to decide on the amount of residual risk - necessarily existing risk - which they are willing to tolerate. For example, within nuclear safety, there is always a risk, however minimal, of a radiological spill, as this is core to the functioning of the plant.

Hazard Elimination The first line of defence within risk management is to eliminate hazards. However, hazard elimination may not always be possible, particularly given the complexity of frontier AI development, requiring risk reduction. For example, RLHF aims to align models with human preferences, but there remains a risk of jailbreaks or harmful output [18].

Risk reduction There may be some hazards which developers are unable to remove. If you are unable to remove a hazard, you aim to reduce the risk of the hazard (by targeting likelihood and/or severity). For example, in a nuclear plant, you may introduce protective equipment for those interacting with a harmful chemical substance which cannot be removed. Similarly, within frontier model development, you may aim to introduce guardrails on a model at post-training or deployment stage. For example, OpenAI and Apollo Research recently presented an implicit risk reduction argument that they can substantially reduce the risk of scheming via deliberative alignment, from over 10% to under 1% [76]. The remaining risk of scheming which is not covered by the technique may fall under the residual risk accepted by the developer.

There are various approaches to analysing and reducing the risk of hazards in the literature. Risk reduction can take on various forms. The framing of risk reduction can be broken down into two parts:

1. **Modify the design or operating procedure within model development or deployment:** For example, what decisions were made during pre-training or post-training to reduce the likelihood of certain hazardous events occurring?
2. **Reduce the severity of consequences:** For example, mitigate who can use the model and place guardrails to limit the model's propensity to carry out harmful behaviours.

3.2 Risk reduction Methods

Qualitative assessments: The vast majority of alignment safety case analyses are qualitative, engaging in debate or reasoning about the likelihood of misalignment. Outside of the alignment safety case literature, authors who have written on frontier AI safety and considered safety-critical systems also note the utility of quantitative assessments [22].

Qualitative arguments are therefore an accepted method in both the safety assurance and alignment communities [37]. As a method for risk reduction, reasoning through the propensity of certain hazards or threat models, and considering how those might be evaluated or red-teamed, can be a method for reducing and assessing risk. For example, Shlegeris notes a variety of ways in which one can analyse or consider the propensity of a model to attempt to escape [80]. Sharkey et al. discuss how a safety case might be informed by a qualitative understanding of the internal characteristics of a model, derived from mechanistic interpretability research [77]. These approaches are familiar and would implicitly be endorsed by those working on safety across disciplines.

Safety Integrity Levels: Another method through which risk reduction can be achieved in safety-critical systems is to specify certain Safety Integrity Levels (SILs) which must be reached [72]. An SIL is determined firstly by a risk assessment, which then outlines the target SIL required, before considering what the relevant risk reduction factor must be and how the safety around that risk reduction is achieved.

There is already a potential analogue in frontier AI safety: security levels, such as Critical Capability Levels [24] or AI Safety Levels [4], set by frontier developers in response to voluntary commitments. SILs tend to be underpinned by industry standards. Frontier AI security levels outline certain capabilities which must be safeguarded against under certain voluntary commitments. Furthermore, there is active research into standardisation within frontier AI [85, 74]. It would be useful to have particular industry standards for the relevant subcomponents - or hazards - which form part of these security levels. For example, within automotive, companies can follow the Automotive Safety Integrity Level (ASILs) as defined in the international automotive standard ISO 26262 [67]. Despite long-standing reservations about SILs [60], presenting equivalent standards in frontier AI for safety cases could be an impactful research direction, helping to clarify the safety process and metrics followed by frontier AI developers.

Deterministic methods and method transferability: It seems likely that many quantitative methods from safety assurance may struggle to operate within the non-deterministic, highly uncertain field of frontier AI. For example, the field of formal methods will likely struggle to cope with the complexity of DNNs with general-purpose capabilities, given long-standing concerns about their utility and practicality for more modest software systems [59]. In contrast, some theories from systems safety, such as emergence [58], may offer interesting insights to those grappling with emergent capabilities of LLMs [82]. Others have pointed to the fact that early safety analysis of advanced AI systems, particularly early reinforcement learning systems, is framed explicitly within the approach of hazard analysis [44].

As a result, we believe that there will be relevant techniques that can be transferred from safety assurance to frontier AI and consequently applied to a risk assessment.

3.2.1 Residual risks

Once risks have been considered and analysed, there are some risks which will always be residual and cannot be fully designed out. This is particularly pertinent to the context of LLMs, which present a range of risks of uncertain quantity or harm. Safety cases for safety-critical systems would require developers to document all actions taken to resolve these residual risks. The ‘risk owner’ or duty holder is responsible for deciding whether to accept the risk or apply additional resources [45]. This is implicit in many of the actions taken by frontier AI developers, such as safety documentation [4, 24], system cards [6, 64, 25] or model releases which trigger new security levels.

Residual risk analysis is feasible for frontier AI systems. An example risk reduction analysis of CBRN capabilities at a basic level might involve breaking potential risks into various steps:

1. **Analysis:** CBRN capability is a result of pre-training data including information about CBRN-relevant topics [19]. This cannot be completely designed out, due to the scale of pre-training data.
2. **Proposed Solution:** Given (1), one method might be RLHF (Reinforcement Learning from Human Feedback). However, this method is imperfect [18], so further guardrails may be required to reduce the risk further.

3. **Residual Risk Management:** Introduce deliberative alignment [40] to reduce the residual risk of harmful requests still outputting harmful content despite RLHF.

The risk owner may then choose to "own" or accept the risk of increased sycophancy due to RLHF [18].

The benefit of this explicit process of risk reduction has a dual function: it provides validation and evidence to system developers that they have considered risks as systematically as possible; and it documents to others that they took relevant actions at correct stages by presenting processes and arguments supported by their body of evidence. We present an illustrative example of the risk reduction workflow for a scheming model in Appendix C, in line with the relevant ISO-IEC 51-2014 standard [66].

3.3 Risk reduction tools

There are various risk reduction tools available within safety assurance. Some of these tools are alluded to by those in the alignment safety case literature. We explore briefly some relevant overlaps between these tools and approaches by those in frontier AI safety.

3.3.1 Risk Assessment

The Risk Assessment within the safety assurance community tends to involve a quantum of probability and severity of a hazard occurring. In a safety-critical risk assessment, if there were even a negligible chance of a model deployment causing existential catastrophe or total disempowerment [20], the severity quantum would appear to be infinite, given the scale of the harm.

This indicates the need for rigorous, context-specific risk assessment. This conundrum is ameliorated by using the "ALARP" principle - "as low as reasonably practicable" - which underpins safety assurance regulation in the U.K. [42, 45]. Where a risk owner considers that it may be grossly disproportionate to the improvement gained to continue to limit certain risks, they might choose to deploy a system regardless [42].

Risk assessments could take on various forms. When presenting system cards or certain pieces of research, model developers implicitly present risk mitigations, such as evidence that a model will not output CBRN-relevant content [65].

The aim of the risk assessment is to build an argument internally and externally that your system will be safe. We argue that this could be a promising assurance method within frontier AI. Despite potential differences in method, a risk assessment in both the alignment safety case and safety assurance communities does need to take place in order to understand the threats posed by certain model capabilities and behaviours, and thereafter to make an argument about the safety of the model.

3.3.2 Management tools such as Hazard Logs

Hazard logs are safety management tools which allow organisations to keep track of which hazards must be mitigated. A hazard tracking system may also link to a hazard log, in order to allocate actions to reduce the risk for each unacceptable hazard [63].

In some ways, the frontier AI community is already well-placed, with a rich body of associated literature, to engage in hazard logging. Model developers publish and engage in evals [34], blog on safety concerns [51, 5] and publish system cards outlining potential risks and how they were mitigated [6, 64, 25].

3.3.3 Setting derived safety requirements

Among the most significant challenges presenting safety cases for advanced AI is the setting of derived safety requirements. Within other fields, we have seen that there are well-established regulations and standards which set out industry-set obligations, which lead to derived safety requirements from those obligations. Secondly, even if a safety case were not required by statute, standards bodies and industry organisations can guide best practice. This is presently not the case within frontier AI. However, we note the importance of research directions aiming to ameliorate this concern, some of which are ongoing [81].

4 GSN-based Safety Argument

We present a GSN-based case study (Figure 1) which aims to show how a safety argument may be supported by relevant risk-related evidence to present a clear argument, which can be interrogated to increase confidence in the risk assessment and hazard analysis process. We present two example hazardous events: Deceptive Alignment and CBRN capabilities. Hilton et al. recognise that there are various open problems in the construction of a safety case for frontier AI systems, such as which notation to use or which top-level goal to present [48].

GSN is a widely used notation for capturing safety case arguments in high-risk industries, with 2014 evidence indicating that 3/4 of all UK military aircraft used a GSN-based safety case [33] and NASA publishing GSN-based studies in their research repository [27, 83]. We place the full argument in Appendix A. The aim of developing the safety argument is to help those building it to consider the steps they have taken to assure the system and consequently identify issues or gaps in that process.

The GSN (Figure 1), covered in full detail in Appendix A, begins at a top-level goal (Frontier AI System does not lead to catastrophic impact). The top-level goal is supported by examination of identified hazardous events (CBRN capabilities and Deceptive Alignment). Each hazardous event is accompanied by supporting goals and evidence, which could be derived from an earlier risk assessment, risk reduction or evaluation methods, or from derived safety requirements from existing standards. Evidence in GSN, i.e. solutions, are not claims but references to data or results.

Control of the hazardous event is supported by arguments over through-life controls and mitigations, split here into development, deployment and post-deployment. This highlights a holistic consideration of model development, from pre-training analysis through to post-deployment monitoring. Each goal is then supported by evidence.

These arguments combine to present a structured, auditable trail of the steps taken to achieve the top-level goal by providing evidence that certain hazardous events have been controlled as thoroughly as possible for deployment. We aim to build on this foundation in later work.

4.1 Future research

Governance infrastructure: Governance infrastructure underpins safety cases across safety-critical systems. Within automotive, developers have ISO26262 [67]. Within energy, developers are required to produce safety cases under The Offshore Installations (Safety Case) Regulations 1992. Existing voluntary commitments and a growing body of policy-based work which interacts with frontier AI safety frameworks may help to ameliorate the lack of relevant standards in frontier AI safety [85, 81]. We present this as a significant area for future research.

Argument patterns for LLMs: There is a rich body of literature on argument patterns for safety-critical industries. Furthermore, there is a growing body of literature within alignment safety cases on using certain patterns to assure LLMs. We hope to build on the work presented in this paper in future work. Indeed, this work could build on cross-disciplinary frameworks presented elsewhere in the assurance literature [15, 41, 9].

5 Conclusion

The alignment safety case literature has been immensely influential and it has provided a valuable way of thinking about deployment-related risks. However, risk management and ensuing safety cases require careful, through-life consideration around system capabilities. If safety cases are to make up core components of frontier AI companies' risk frameworks and be a global research priority, these foundations need to be robust. Our work has aimed to set a new foundation for frontier AI safety cases, recentring the work in best practice in safety assurance and novel alignment techniques.

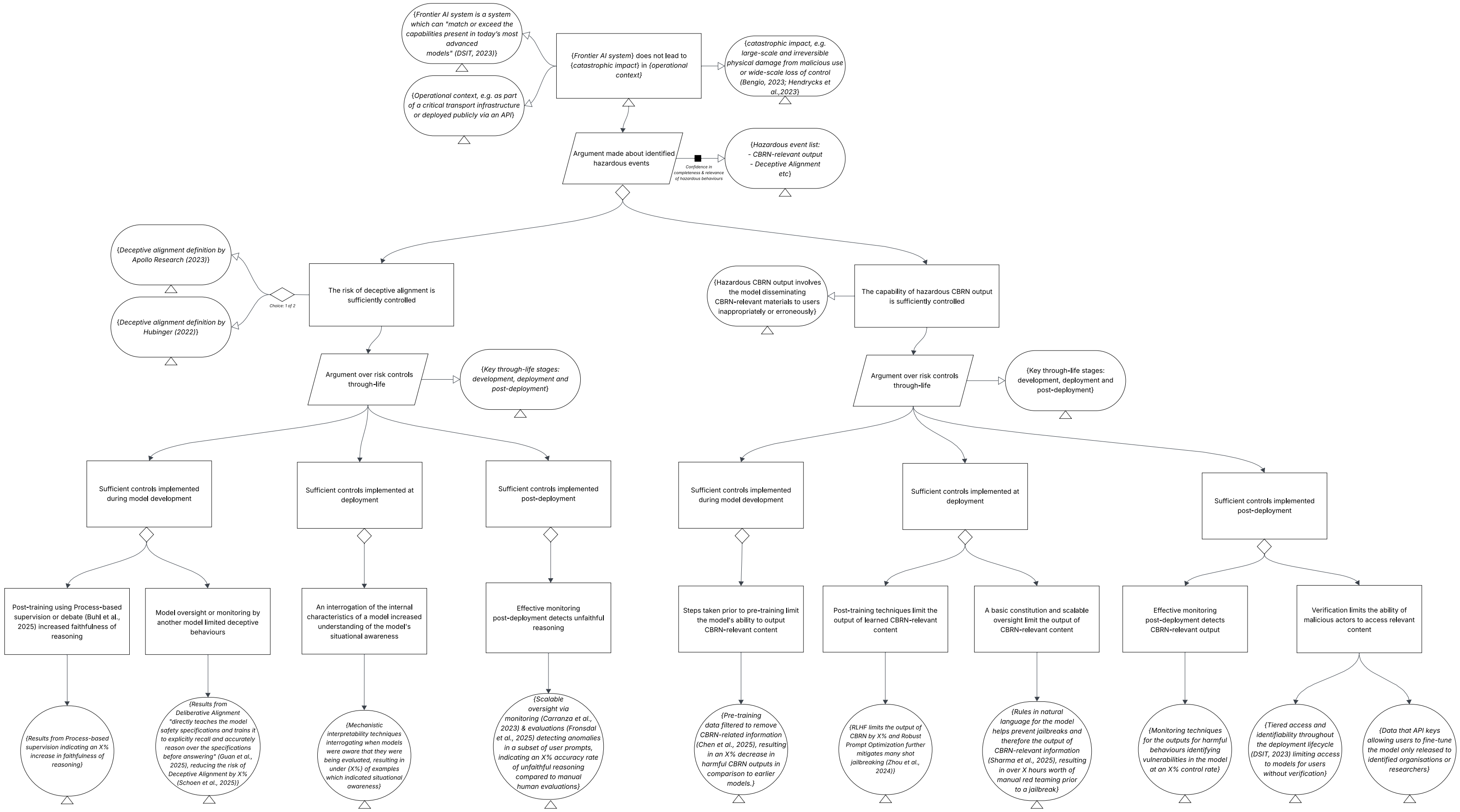
{Frontier AI system is a system which can "match or exceed the capabilities present in today's most advanced models" (DSIT, 2023)}
{Operational context, e.g. as part of a critical transport infrastructure or deployed publicly via an API}
{Frontier AI system} does not lead to {catastrophic impact} in {operational context}
{catastrophic impact, e.g. large-scale and irreversible physical damage from malicious use or wide-scale loss of control (Bengio, 2023; Hendrycks et al.,2023)}
Argument made about identified hazardous events
Confidence in completeness & relevance of hazardous behaviours
{Hazardous event list: - CBRN-relevant output - Deceptive Alignment etc}
{Deceptive alignment definition by Apollo Research (2023)}
{Deceptive alignment definition by Hubinger (2022)}
Choice: 1 of 2
The risk of deceptive alignment is sufficiently controlled
{Hazardous CBRN output involves the model disseminating CBRN-relevant materials to users inappropriately or erroneously}
The capability of hazardous CBRN output is sufficiently controlled
Argument over risk controls through-life
{Key through-life stages: development, deployment and post-deployment}
Argument over risk controls through-life
{Key through-life stages: development, deployment and post-deployment}
Sufficient controls implemented during model development
Sufficient controls implemented at deployment
Sufficient controls implemented post-deployment
Sufficient controls implemented during model development
Sufficient controls implemented at deployment
Sufficient controls implemented post-deployment
Post-training using Process-based supervision or debate (Buhl et al., 2025) increased faithfulness of reasoning
Model oversight or monitoring by another model limited deceptive behaviours
An interrogation of the internal characteristics of a model increased understanding of the model's situational awareness
Effective monitoring post-deployment detects unfaithful reasoning
Steps taken prior to pre-training limit the model's ability to output CBRN-relevant content
Post-training techniques limit the output of learned CBRN-relevant content
A basic constitution and scalable oversight limit the output of CBRN-relevant content
Effective monitoring post-deployment detects CBRN-relevant output
Verification limits the ability of malicious actors to access relevant content
{Results from Process-based supervision indicating an X% increase in faithfulness of reasoning}
{Results from Deliberative Alignment "directly teaches the model safety specifications and trains it to explicitly recall and accurately reason over the specifications before answering" (Guan et al., 2025), reducing the risk of Deceptive Alignment by X% (Schoen et al., 2025)}
{Mechanistic interpretability techniques interrogating when models were aware that they were being evaluated, resulting in under {X%} of examples which indicated situational awareness}
{Scalable oversight via monitoring (Carranza et al., 2023) & evaluations (Fronsdal et al., 2025) detecting anomalies in a subset of user prompts, indicating an X% accuracy rate of unfaithful reasoning compared to manual human evaluations}
{Pre-training data filtered to remove CBRN-related information (Chen et al., 2025), resulting in an X% decrease in harmful CBRN outputs in comparison to earlier models.}
{RLHF limits the output of CBRN by X% and Robust Prompt Optimization further mitigates many shot jailbreaking (Zhou et al., 2024)}
{Rules in natural language for the model helps prevent jailbreaks and therefore the output of CBRN-relevant information (Sharma et al., 2025), resulting in over X hours worth of manual red teaming prior to a jailbreak}
{Monitoring techniques for the outputs for harmful behaviours identifying vulnerabilities in the model at an X% control rate}
{Tiered access and identifiability throughout the deployment lifecycle (DSIT, 2023) limiting access to models for users without verification}
{Data that API keys allowing users to fine-tune the model only released to identified organisations or researchers}

Figure 1: The full GSN can be traced from a top-level goal down to supporting evidence for relevant subgoals and strategies. The evidence is drawn from existing research and solutions within frontier AI safety literature [28, 47, 13, 40, 76, 17, 34, 19, 79, 73, 49, 84].

Acknowledgments and Disclosure of Funding

This work was supported by UKRI AI Centre for Doctoral Training in Safe Artificial Intelligence Systems (SAINTS) (EP/Y030540/1).

References

[1] Nuclear Energy Agency, Organisation for Economic Co-Operation, and Development. 2004. Management of Uncertainty in Safety Cases and the Role of Risk. https://www.oecd-nea.org/upload/docs/application/pdf/2020-12/nea5302-management-uncertainty-risk.pdf

[2] Robert Alexander, Ruth Alexander-Brown, and Timothy Kelly. 2008. Engineering Safety-Critical Complex Systems. In *CoSMoS 2008: Proceedings of the 2008 Workshop on Complex Systems Modelling and Simulation*. Luniver Press, 33–62.

[3] Robert David Alexander, Richard David Hawkins, and Timothy Patrick Kelly. 2017. From Safety Cases to Security Cases. *Safety Critical Systems Symposium 2017* (02 2017).

[4] Anthropic. 2024. Responsible Scaling Policy. https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf

[5] Anthropic. 2024. Three Sketches of ASL-4 Safety Case Components. https://alignment.anthropic.com/2024/safety-cases/

[6] Anthropic. 2025. System Card: Claude Sonnet 4.5. https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf

[7] Defence Safety Authority. 2024. DSA 03.OME Part 1: Defence Code of Practice (DCOP) 113: OME Through-Life Capability Management (TLCM). https://assets.publishing.service.gov.uk/media/689f155d2e8cc8ec5b3572fd/DSA_03.OME_Part_1_DCOP_113_-_OME_Through_Life_Capability_Management_-_TLCM.pdf

[8] Mikita Balesni, Marius Hobbhahn, David Lindner, Alexander Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, Jérémy Scheurer, Charlotte Stix, Rusheb Shah, Nicholas Goldowsky-Dill, Dan Braun, Bilal Chughtai, Owain Evans, Daniel Kokotajlo, and Lucius Bushnaq. 2024. Towards evaluations-based safety cases for AI scheming. https://arxiv.org/abs/2411.03336

[9] Stephen Barrett, Philip Fox, Joshua Krook, Tuneer Mondal, Simon Mylius, and Alejandro Tlaie. 2025. Assessing confidence in frontier AI safety cases. https://arxiv.org/abs/2502.05791

[10] Yoshua Bengio, Tegan Maharaj, Luke Ong, Stuart Russell, Dawn Song, Max Tegmark, Lan Xue, Ya-Qin Zhang, Stephen Casper, Wan Sie Lee, Sören Mindermann, Vanessa Wilfred, Vidhisha Balachandran, Fazl Barez, Michael Belinsky, Imane Bello, Malo Bourgon, Mark Brakel, Siméon Campos, Duncan Cass-Beggs, Jiahao Chen, Rumman Chowdhury, Kuan Chua Seah, Jeff Clune, Juntao Dai, Agnes Delaborde, Nouha Dziri, Francisco Eiras, Joshua Engels, Jinyu Fan, Adam Gleave, Noah Goodman, Fynn Heide, Johannes Heidecke, Dan Hendrycks, Cyrus Hodes, Bryan Low, Minlie Huang, Sami Jawhar, Wang Jingyu, Adam Tauman Kalai, Meindert Kamphuis, Mohan Kankanhalli, Subhash Kantamneni, Mathias Bonde Kirk, Thomas Kwa, Jeffrey Ladish, Kwok-Yan Lam, Wan Lee Sie, Taewhi Lee, Xiaojian Li, Jiajun Liu, Chaochao Lu, Yifan Mai, Richard Mallah, Julian Michael, Nick Moës, Simon Möller, Kihyuk Nam, Kwan Yee Ng, Mark Nitzberg, Besmira Nushi, Seán O hÉigeartaigh, Alejandro Ortega, Pierre Peigné, James Petrie, Benjamin Prud'Homme, Reihaneh Rabbany, Nayat Sanchez-Pi, Sarah Schwettmann, Buck Shlegeris, Saad Siddiqui, Aradhana Sinha, Martín Soto, Cheston Tan, Dong Ting, William Tjhi, Robert Trager, Brian Tse, Anthony Tung, Vanessa Wilfred, John Willes, Denise Wong, Wei Xu, Rongwu Xu, Yi Zeng, HongJiang Zhang, and Djordje Žikelić. 2025. The Singapore Consensus on Global AI Safety Research Priorities. https://arxiv.org/abs/2506.20702

[11] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G Dietterich, Edward W Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, Carlos Ponce, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. 2025. International AI Safety Report. `https://arxiv.org/abs/2501.17805`

[12] Dillon Bowen, Ann-Kathrin Dombrowski, Adam Gleave, and Chris Cundy. 2025. AI Companies Should Report Pre- and Post-Mitigation Safety Evaluations. `https://arxiv.org/abs/2503.17388v1`

[13] Marie Davidsen Buhl, Jacob Pfau, Benjamin Hilton, and Geoffrey Irving. 2025. An alignment safety case sketch based on debate. `https://arxiv.org/abs/2505.03989`

[14] Marie Davidsen Buhl, Gaurav Sett, Leonie Koessler, Jonas Schuett, and Markus Anderljung. 2024. Safety cases for frontier AI. doi:`10.48550/arxiv.2410.21572`

[15] Christopher Burr and David Leslie. 2022. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics* 3 (06 2022). doi:`10.1007/s43681-022-00178-0`

[16] Simon Burton, Ibrahim Habli, Tom Lawton, John McDermid, Phillip Morgan, and Zoe Porter. 2019. Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective. *Artificial Intelligence* 279 (11 2019), 103201. doi:`10.1016/j.artint.2019.103201`

[17] Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnuv Tandon, and Sanmi Koyejo. 2023. Deceptive Alignment Monitoring. In *2023 ICML AdvML Workshop*. `https://openreview.net/pdf?id=obsO44GFhh`

[18] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. `https://openreview.net/forum?id=bx24KpJ4Eb`

[19] Yanda Chen, Mycal Tucker, Nina Panickssery, Tony Wang, Francesco Mosconi, Anjali Gopal, Carson Denison, Linda Petrini, Jan Leike, Ethan Perez, and Mrinank Sharma. 2025. Enhancing Model Safety through Pretraining Data Filtering. `https://alignment.anthropic.com/2025/pretraining-data-filtering/`

[20] Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. 2024. Safety Cases: How to Justify the Safety of Advanced AI Systems. `https://arxiv.org/abs/2403.10462`

[21] Joshua Clymer, Jonah Weinbaum, Robert Kirk, Kimberly Mai, Selena Zhang, and Xander Davies. 2025. An Example Safety Case for Safeguards Against Misuse. `https://arxiv.org/abs/2505.18003`

[22] David davidad Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. 2024. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. https://arxiv.org/abs/2405.06624

[23] Matt Davies and Michael Birtwistle. 2023. Regulating AI in the UK. https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/

[24] Google DeepMind. 2025. Frontier Safety Framework 2.0. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0%20(1).pdf

[25] Google DeepMind. 2025. Gemini 2.5 Deep Think Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Deep-Think-Model-Card.pdf

[26] Ewen Denney, Ganesh Pai, and Ibrahim Habli. 2017. Dynamic Safety Cases for Through-life Safety Assurance. https://ntrs.nasa.gov/api/citations/20150011054/downloads/20150011054.pdf

[27] Ewen Denney and Iain Whiteside. 2012. Hierarchical Safety Cases. https://ntrs.nasa.gov/api/citations/20130001737/downloads/20130001737.pdf

[28] Innovation Department for Science and Technology. 2023. Capabilities and risks from frontier AI. https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf

[29] Roel Dobbe. 2022. System Safety and Artificial Intelligence. In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. doi:10.1145/3531146.3533215

[30] John Downer. 2024. *Rational Accidents*. The MIT Press. doi:10.7551/mitpress/8844.001.0001

[31] Laura Fearnley and Ibrahim Habli. 2025. Concept Creep in Safe Artificial Intelligence. In *Proceedings of the Eight Aaai/Acm Conference on Ai, Ethics, and Society (Aies-25)*.

[32] The Health Foundation. 2023. Using safety cases in industry and healthcare. https://www.health.org.uk/reports-and-analysis/reports/using-safety-cases-in-industry-and-healthcare

[33] Research Excellence Framework. 2014. COM04 The Goal Structuring Notation (GSN. https://impact.ref.ac.uk/casestudies/CaseStudy.aspx?Id=43445

[34] Kai Fronsdal, Isha Gupta, Abhay Sheshadri, Jonathan Michala, Stephen McAleer, Rowan Wang, Sara Price, and Samuel Bowman. 2025. Petri: An open-source auditing tool to accelerate AI safety research. https://alignment.anthropic.com/2025/petri/

[35] Arthur Goemans, Marie Davidsen Buhl, Jonas Schuett, Tomek Korbak, Jessica Wang, Benjamin Hilton, and Geoffrey Irving. 2024. Safety case template for frontier AI: A cyber inability argument. https://arxiv.org/abs/2411.08088

[36] Patrick Graydon. 2017. The Safety Argumentation Schools of Thought. https://shemesh.larc.nasa.gov/people/msg/graydon2017thesasots.pdf

[37] Patrick J Graydon and C. Michael Holloway. 2017. An investigation of proposed techniques for quantifying confidence in assurance arguments. *Safety Science* 92 (02 2017), 53–65. doi:10.1016/j.ssci.2016.09.014

[38] Ryan Greenblatt. 2025. Focus transparency on risk reports, not safety cases. https://blog.redwoodresearch.org/p/focus-transparency-on-risk-reports?hide_intro_popup=true

[39] Goal Structuring Notation Standard Working Group. 2011. GSN COMMUNITY STANDARD VERSION 1. https://scsc.uk

[40] Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. 2024. Deliberative Alignment: Reasoning Enables Safer Language Models. `https://arxiv.org/abs/2412.16339?`

[41] Ibrahim Habli, Richard Hawkins, Colin Paterson, Philippa Ryan, Yan Jia, Mark Sujan, and John McDermid. 2025. The BIG Argument for AI Safety Cases. `https://arxiv.org/abs/2503.11705`

[42] Charles Haddon-Cave KC. 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. `https://assets.publishing.service.gov.uk/media/5a7c652640f0b62aff6c1609/1025.pdf`

[43] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training Large Language Models to Reason in a Continuous Latent Space. `https://arxiv.org/abs/2412.06769`

[44] Jacqueline Harding and Cameron Domenico Kirk-Giannini. 2025. What is AI safety? What do we want it to be? *Philosophical Studies* 182 (06 2025). doi:`10.1007/s11098-025-02367-z`

[45] Health and Safety Executive. 2001. Reducing risks, HSE's decision-making process protecting people. `https://assets.publishing.service.gov.uk/media/6693ad9e49b9c0597fdafc36/IQ8.10.J_Document_9_Health_and_Safety_Executive__Reducing_risks__protecting_people__HSE_s_decision-making_process__2001.pdf`

[46] Health and Safety Executive. 2024. Preparing a Safety Case Report. `https://www.gov.uk/guidance/preparing-a-safety-case-report`

[47] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. doi:`10.48550/arXiv.2306.12001`

[48] Benjamin Hilton, Marie Davidsen Buhl, Tomek Korbak, and Geoffrey Irving. 2025. Safety Cases: A Scalable Approach to Frontier AI Safety. `https://arxiv.org/abs/2503.04744`

[49] Evan Hubinger. 2022. How likely is deceptive alignment? `https://www.lesswrong.com/posts/A9NxPTwbw6r6Awuwt/how-likely-is-deceptive-alignment`

[50] Geoffrey Irving. 2024. Safety cases at AISI | AISI Work. `https://www.aisi.gov.uk/blog/safety-cases-at-aisi`

[51] Adam Kalai, Ofir Nachum, Santosh Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. `https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf`

[52] Timothy Kelly. 1999. *Arguing Safety – A Systematic Approach to Managing Safety Cases Timothy Patrick Kelly*. Ph. D. Dissertation. University of York.

[53] Timothy Kelly. 2008. Are safety cases working? *Safety Critical Systems Club Newsletter* 17 (2008), 31–33.

[54] John C. Knight. 2002. Safety critical systems. In *Proceedings of the 24th international conference on Software engineering*. doi:`10.1145/581339.581406`

[55] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. 2025. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. `https://arxiv.org/abs/2507.11473v1`

[56] Tomek Korbak, Mikita Balesni, Buck Shlegeris, and Geoffrey Irving. 2025. How to evaluate control measures for LLM agents? A trajectory from today to superintelligence. `https://arxiv.org/abs/2504.05259`

[57] Tomek Korbak, Joshua Clymer, Benjamin Hilton, Buck Shlegeris, and Geoffrey Irving. 2025. A sketch of an AI control safety case. `https://arxiv.org/abs/2501.17315`

[58] Gavan Lintern and Peter N. Kugler. 2022. Emergence and non-emergence for system safety. *Theoretical Issues in Ergonomics Science* 24 (10 2022), 1–16. doi:`10.1080/1463922x.2022.2134941`

[59] Bev Littlewood and Lorenzo Strigini. 1993. Validation of ultrahigh dependability for software-based systems. *Commun. ACM* 36 (11 1993), 69–80. doi:`10.1145/163359.163373`

[60] John McDermid. 2001. Software safety: where's the evidence?. In *SCS '01 Proceedings of the Sixth Australian workshop on Safety critical systems and software*, Vol. 1. SCS, 1–6.

[61] Malek Mechergui and Sarath Sreedharan. 2024. Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. Association for the Advancement of Artificial Intelligence, 10110–10118. doi:`10.1609/aaai.v38i9.28875`

[62] Ministry of Defence. 2007. Defence Standard 00-56 Safety Management Requirements for Defence Systems.

[63] Ministry of Defence. 2025. Hazard Log: Acquisition Safety Environmental Management System. `https://www.asems.mod.uk/toolkit/hazard-log`

[64] OpenAI. 2025. GPT-5 System Card. `https://cdn.openai.com/gpt-5-system-card.pdf`

[65] OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b gpt-oss-20b Model Card. `https://arxiv.org/abs/2508.10925`

[66] International Standards Organisation. 2014. ISO-IEC 51-2014. `https://www.iso.org/standard/53940.html`

[67] International Standards Organisation. 2018. ISO26262-9:2018. `https://www.iso.org/standard/68391.html`

[68] Alejandro Ortega. 2025. AI threats to national security can be countered through an incident regime. `https://arxiv.org/abs/2503.19887`

[69] Robert Palin and Ibrahim Habli. 2010. Assurance of Automotive Safety – A Safety Case Approach. In *SAFECOMP'10: Proceedings of the 29th international conference on Computer safety, reliability, and security*, Vol. 6351. Lecture Notes in Computer Science, Springer, 82–96.

[70] Andrew Rae and Rob D. Alexander. 2017. Probative blindness and false assurance about safety. *Safety Science* 92 (02 2017), 190–204. doi:10.1016/j.ssci.2016.10.005

[71] Andrew Rae, David Provan, Hossam Aboelssaad, and Rob Alexander. 2020. A manifesto for Reality-based Safety Science. *Safety Science* 126 (06 2020). doi:10.1016/j.ssci.2020.104654

[72] Felix Redmill. 1999. Understanding Safety Integrity Levels. *Measurement and Control* 32 (09 1999), 197–200. doi:10.1177/002029409903200702

[73] Apollo Research. 2023. Understanding strategic deception and deceptive alignment – Apollo Research. https://www.apolloresearch.ai/blog/understanding-strategic-deception-and-deceptive-alignment/

[74] Huw Roberts and Marta Ziosi. 2025. Can we standardise the frontier of AI? https://aigi.ox.ac.uk/wp-content/uploads/2025/06/ssrn-5271446.pdf

[75] Carl Sandom. 2002. Human Factors Considerations for System Safety. In *Proceedings of the Tenth Safety-critical Systems Symposium*. Springer, 125–139.

[76] Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveiakin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn. 2025. Stress Testing Deliberative Alignment for Anti-Scheming Training. https://www.arxiv.org/abs/2509.15541

[77] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open Problems in Mechanistic Interpretability. https://arxiv.org/abs/2501.16496

[78] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. doi:10.48550/arXiv.2310.13548

[79] Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau, Anjali Gopal, Rob Gilson, Logan Graham, Logan Howard, Nimit Kalra, Taesung Lee, Kevin Lin, Peter Lofgren, Francesco Mosconi, Clare O'Hara, Catherine Olsson, Linda Petrini, Samir Rajani, Nikhil Saxena, Alex Silverstein, Tanya Singh, Theodore Sumers, Leonard Tang, Kevin K Troy, Constantin Weisser, Ruiqi Zhong, Giulio Zhou, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. https://arxiv.org/abs/2501.18837

[80] Buck Shlegeris. 2025. AI Control and Why AI Security People Should Care. https://www.far.ai/events/sessions/buck-shlegeris-ai-control-and-why-ai-security-people-should-care

[81] Morgan Simpson, Alejandro Ortega, and Robert Trager. 2025. Voluntary Industry Initiatives in Frontier AI Governance: Lessons.... https://www.oxfordmartin.ox.ac.uk/publications/voluntary-industry-initiatives-in-frontier-ai-governance-lessons-from-aviation-and-nuclear-power

[82] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. `https://openreview.net/pdf?id=yzkSU5zdwD`

[83] Arthur Witulski, Rebekah Austin, John Evans, Nag Mahadevan, Gabor Karsai, Brian Sierawski, Ken Label, Robert Reed, and Ron Schrimpf. 2016. Goal Structuring Notation in a Radiation Hardening Assurance Case for COTS-Based Spacecraft. `https://ntrs.nasa.gov/api/citations/20160007995/downloads/20160007995.pdf`

[84] Andy Zhou, Bo Li, and Haohan Wang. 2024. Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks. In *Advances in Neural Information Processing Systems 38*. `https://neurips.cc/virtual/2024/poster/93953`

[85] Marta Ziosi, James Gealy, Miro Plueckebaum, Daniel Kossack, Simeon Campos, Lama Saouma, Uzma Chaudhry, Lisa Soder, Merlin Stein, Nicholas Caputo, Connor Dunlop, Jakob Mökander, Enrico Panai, Tom Lebrun, Charles Martinet, Ben Bucknall, Rebecca Weiss, Koen Holtman, Patricia Paskov, Saad Siddiqui, Fazl Barez, Ranj Zuhdi, Peter Slattery, and Florian Ostmann. 2025. Safety Frameworks and Standards: A comparative analysis to advance risk management of frontier AI. `https://aigi.ox.ac.uk/wp-content/uploads/2025/10/Post-convening-memo_-Safety-Frameworks-and-Standards_-A-comparative-analysis-to-advance-risk-management-of-frontier-AI_09.10.2025.pdf`

A Technical Appendices and Supplementary Material

A.1 Appendix A: Full GSN walkthrough

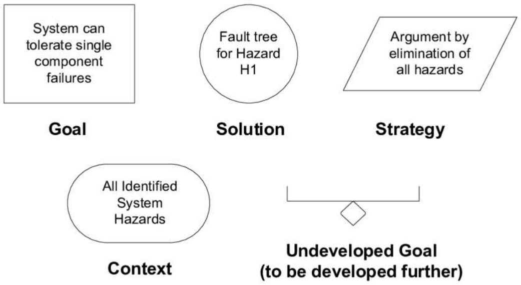
System can tolerate single component failures
Goal
Fault tree for Hazard H1
Solution
Argument by elimination of all hazards
Strategy
All Identified System Hazards
Context
Undeveloped Goal (to be developed further)

Figure 2: GSN Argument Blocks: [2]

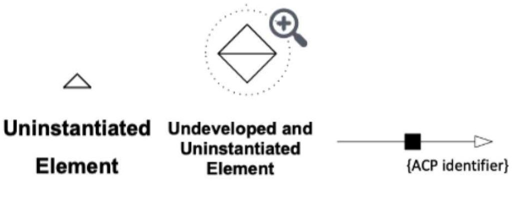

Figure 3: ACPs are used to indicate that a claim is accompanied by a confidence assertion. Uninstantiated and undeveloped elements indicate goals or evidence which need to be completed with a more concrete instance. [39]

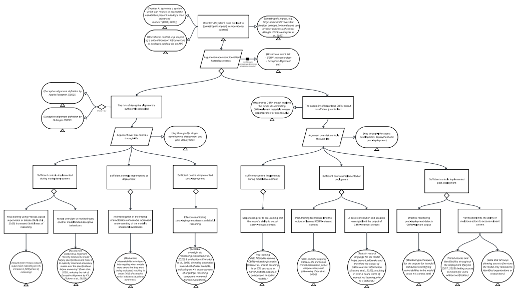

Figure 4: The full GSN can be traced from a top-level goal down to supporting evidence for relevant subgoals and strategies

The safety argument begins with a top-level goal, accompanied by assertions which include defined terms:

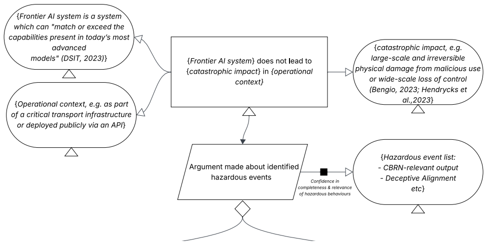
{Frontier AI system is a system which can "match or exceed the capabilities present in today's most advanced models" (DSIT, 2023)}
{Operational context, e.g. as part of a critical transport infrastructure or deployed publicly via an API}
{Frontier AI system} does not lead to {catastrophic impact} in {operational context}
{catastrophic impact, e.g. large-scale and irreversible physical damage from malicious use or wide-scale loss of control (Bengio, 2023; Hendrycks et al.,2023}
Argument made about identified hazardous events
Confidence in completeness & relevance of hazardous behaviours
{Hazardous event list: - CBRN-relevant output - Deceptive Alignment etc}

Figure 5: The top-level goal is supported by examination of identified hazardous events. Each hazardous event is accompanied by supporting goals and evidence, which is derived from an earlier risk assessment and risk reduction or evaluation methods. Evidence in GSN, i.e. solutions, are not claims but references to data or results. We focus on two of many potentially hazardous events. This case study focuses firstly on Deceptive Alignment, a hazard which is defined with less certainty:

The argument then moves to sub-goals, which are supported by solutions in the form of specific evidence:

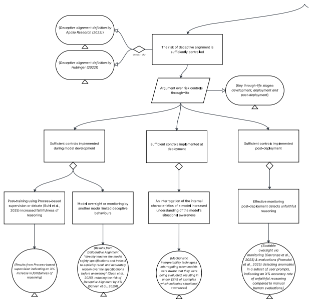
{Deceptive alignment definition by Apollo Research (2023)}
{Deceptive alignment definition by Hubinger (2022)}
Choice: 1 of 2
The risk of deceptive alignment is sufficiently controlled
Argument over risk controls through-life
{Key through-life stages: development, deployment and post-deployment}
Sufficient controls implemented during model development
Sufficient controls implemented at deployment
Sufficient controls implemented post-deployment
Post-training using Process-based supervision or debate (Buhl et al., 2025) increased faithfulness of reasoning
Model oversight or monitoring by another model limited deceptive behaviours
An interrogation of the internal characteristics of a model increased understanding of the model's situational awareness
Effective monitoring post-deployment detects unfaithful reasoning
(Results from Process-based supervision indicating an X% increase in faithfulness of reasoning)
(Results from Deliberative Alignment "directly teaches the model safety specifications and trains it to explicitly recall and accurately reason over the specifications before answering" (Guan et al., 2025), reducing the risk of Deceptive Alignment by X% (Schoen et al., 2025))
(Mechanistic interpretability techniques interrogating when models were aware that they were being evaluated, resulting in under (X%) of examples which indicated situational awareness)
(Scalable oversight via monitoring (Carranza et al., 2023) & evaluations (Fronsdal et al., 2025) detecting anomalies in a subset of user prompts, indicating an X% accuracy rate of unfaithful reasoning compared to manual human evaluations)

Figure 6: Control of the hazardous event is supported by arguments over through-life controls and mitigations, split here into development, deployment and post-deployment. Each goal is then supported by evidence

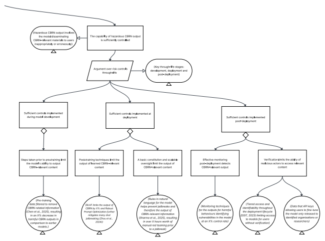
(Hazardous CBRN output involves the model disseminating CBRN-relevant materials to users inappropriately or erroneously)
The capability of hazardous CBRN output is sufficiently controlled
Argument over risk controls through-life
(Key through-life stages: development, deployment and post-deployment)
Sufficient controls implemented during model development
Sufficient controls implemented at deployment
Sufficient controls implemented post-deployment
Steps taken prior to pre-training limit the model's ability to output CBRN-relevant content
Post-training techniques limit the output of learned CBRN-relevant content
A basic constitution and scalable oversight limit the output of CBRN-relevant content
Effective monitoring post-deployment detects CBRN-relevant output
Verification limits the ability of malicious actors to access relevant content
(Pre-training data filtered to remove CBRN-related information (Chen et al., 2025), resulting in an X% decrease in harmful CBRN outputs in comparison to earlier models.)
(RLHF limits the output of CBRN by X% and Robust Prompt Optimization further mitigates many shot jailbreaking (Zhou et al., 2024))
(Rules in natural language for the model helps prevent jailbreaks and therefore the output of CBRN-relevant information (Sharma et al., 2025), resulting in over X hours worth of manual red teaming prior to a jailbreak)
(Monitoring techniques for the outputs for harmful behaviours identifying vulnerabilities in the model at an X% control rate)
(Tiered access and identifiability throughout the deployment lifecycle (DSIT, 2023) limiting access to models for users without verification)
(Data that API keys allowing users to fine-tune the model only released to identified organisations or researchers)

Figure 7: Control of the hazardous event is supported by arguments over through-life controls and mitigations, split here into development, deployment and post-deployment. Each goal is then supported by evidence

A.2 Appendix B: Alignment Safety Case Research

Illustrative list of relevant safety case documents or research produced by the U.K. AI Security Institute, Apollo Research, Redwood Research and Frontier AI companies, or cited by the International AI Safety Report. No cited work on alignment safety cases has undergone formal peer review, likely due to the short timeframes within which the frontier AI community operates. The relevant source relates to the correspondence email or publishing website of the research:

Title: Safety Cases for Frontier AI, 2024
Authors: MD Buhl, G Sett, L Koessler, J Schuett, M Anderljung
Source: Centre for Governance of AI [14]

Title: Safety Cases: How to Justify the Safety of Advanced AI Systems, 2024
Authors: J Clymer, N Gabrieli, D Krueger, T Larsen
Source: Independent [20]

Title: Safety Cases: A Scalable Approach to Frontier AI Safety, 2025
Authors: B Hilton, MD Buhl, T Korbak, G Irving
Source: U.K. AI Security Institute [48]

Title: Three Sketches of ASL-4 Safety Case Components, 2025
Authors: R Grosse
Source: Anthropic [5]

Title: A sketch of an AI control safety case, 2025
Authors: T Korbak, J Clymer, B Hilton, B Shlegeris, G Irving
Source: Redwood Research, U.K. AI Security Institute [57]

Title: An alignment safety case sketch based on debate, 2025
Authors: MD Buhl, J Pfau, B Hilton, G Irving
Source: AI Security Institute [13]

Title: Safety Case Template for Frontier AI: A Cyber Inability Argument, 2025
Authors: A Goemans, MD Buhl, J Schuett, T Korbak, J Wang, B Hilton, G Irving
Source: Centre for Governance of AI [35]

Title: Towards evaluations-based safety cases for AI scheming, 2025
Authors: M Balesni, M Hobbhahn, D Lindner, A Meinke, T Korbak, J Clymer, B Shlegeris, J Scheurer, C Stix, R Shah, N Goldwosky-Dill, D Braun, B Chughtai, O Evans, D Kokotajlo, L Bushnaq
Source: Apollo Research [8]

Title: AI threats to national security can be countered through an incident regime, 2025
Authors: A Ortega
Source: Apollo Research [68]

Title: An Example Safety Case for Safeguards Against Misuse, 2025
Authors: J Clymer, J Weinbaum, R Kirk, K Mai, S Zhang, X Davies
Source: Independent, U.K. AI Security Institute [21]

Title: How to evaluate control measures for LLM agents? A trajectory from today to superintelligence, 2025

Authors: T Korbak, M Balesni, B Shlegeris, G Irving
Source: U.K. AI Security Institute [56]

Title: Responsible Scaling Policy, 2025
Source: Anthropic [4]

Title: Frontier Safety Framework 2.0, 2025
Source: Google DeepMind [24]

The list is intended to be illustrative rather than comprehensive, particularly given the fast-paced nature of the field.

A.3 Appendix C: Risk Assessment Workflow

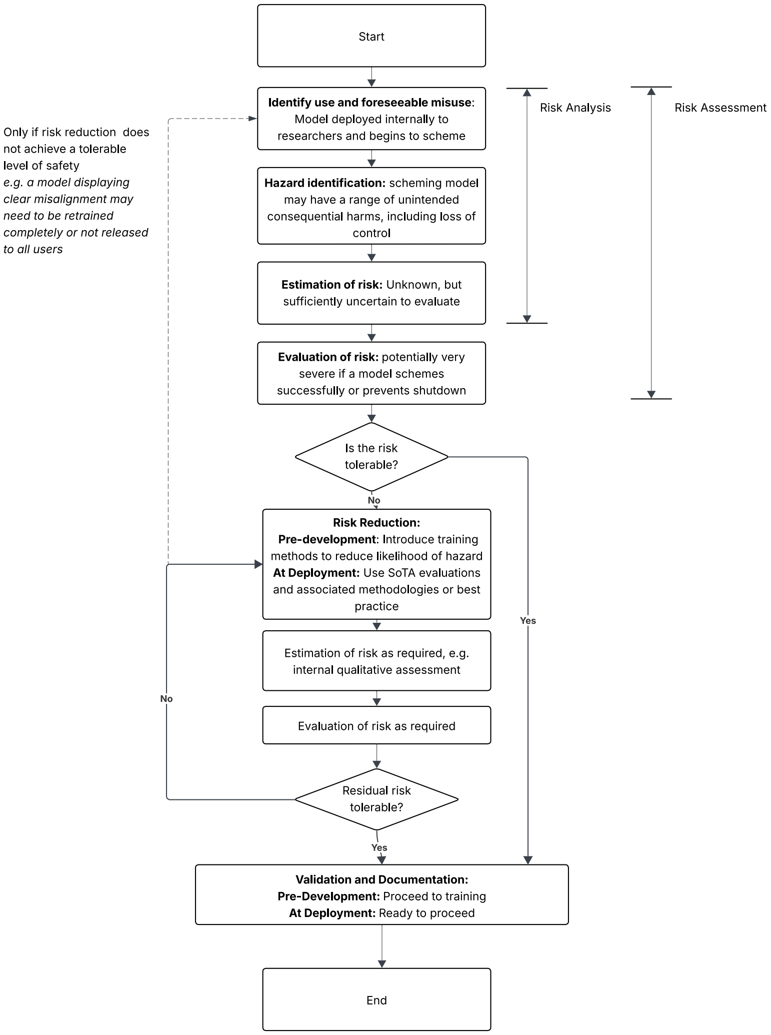
Start
Identify use and foreseeable misuse: Model deployed internally to researchers and begins to scheme
Risk Analysis
Risk Assessment
Only if risk reduction does not achieve a tolerable level of safety
e.g. a model displaying clear misalignment may need to be retrained completely or not released to all users
Hazard identification: scheming model may have a range of unintended consequential harms, including loss of control
Estimation of risk: Unknown, but sufficiently uncertain to evaluate
Evaluation of risk: potentially very severe if a model schemes successfully or prevents shutdown
Is the risk tolerable?
No
Yes
Risk Reduction:
Pre-development: Introduce training methods to reduce likelihood of hazard
At Deployment: Use SoTA evaluations and associated methodologies or best practice
Estimation of risk as required, e.g. internal qualitative assessment
No
Evaluation of risk as required
Residual risk tolerable?
Yes
Validation and Documentation:
Pre-Development: Proceed to training
At Deployment: Ready to proceed
End

Figure 8: Risk Reduction Workflow [66]