The BIG Argument for AI Safety Cases

Ibrahim Habli
Richard Hawkins
Colin Paterson
Philippa Ryan
Yan Jia
Mark Sujan
John McDermid

Centre *for* Assuring Autonomy

The BIG Argument for AI Safety Cases

IBRAHIM HABLI, Centre for Assuring Autonomy, University of York, United Kingdom
RICHARD HAWKINS, Centre for Assuring Autonomy, University of York, United Kingdom
COLIN PATERSON, Centre for Assuring Autonomy, University of York, United Kingdom
PHILIPPA RYAN, Centre for Assuring Autonomy, University of York, United Kingdom
YAN JIA, Centre for Assuring Autonomy, University of York, United Kingdom
MARK SUJAN, Centre for Assuring Autonomy, University of York, United Kingdom
JOHN MCDERMID, Centre for Assuring Autonomy, University of York, United Kingdom

Abstract. We present our Balanced, Integrated and Grounded (BIG) argument for assuring the safety of AI systems. The BIG argument adopts a whole-system approach to constructing a safety case for AI systems of varying capability, autonomy and criticality. Whether the AI capability is narrow and constrained or general-purpose and powered by a frontier or foundational model, the BIG argument insists on a meaningful treatment of safety. It respects long-established safety assurance norms such as sensitivity to context, traceability and risk proportionality. Further, it places a particular focus on the novel hazardous behaviours emerging from the advanced capabilities of frontier AI models and the open contexts in which they are rapidly being deployed. These complex issues are considered within a broader AI safety case that approaches assurance from both technical and sociotechnical perspectives. Examples illustrating the use of the BIG argument are provided throughout the paper.

Additional Key Words and Phrases: AI Safety, Frontier AI, Safety Cases, Assurance.

1 INTRODUCTION

AI is increasingly recognised for its potential to deliver significant benefits at the individual and societal levels, cutting across sectors and national boundaries [1]. AI's use in Healthcare is a classic example [2]. Alleviating pressures on healthcare systems across the world is a global priority. Latest OECD reports highlight AI's capacity to "*automate administrative tasks and free substantial time for healthcare providers to focus on patient care*" [3]. In particular, AI can extend *"health services to remote or underserved areas, improving healthcare access for millions worldwide living hours away from the nearest healthcare facility*"[3].

However, there have been major concerns about the harms, e.g. physical and psychological [4] [5], that the use of AI could cause, especially when the technology is embedded into wider engineered systems and complex social settings [6][7][8]. For instance, accidents and incidents involving autonomous driving have been newsworthy [9]. Two notable examples are an Uber self-driving car crash that led to the death of Elaine Herzberg in Tempe, Arizona in 2018 [10], and the suspending of Cruise 'robotaxi' operations, a subsidiary of General Motors, in San Francisco, California in 2023 following a series of pedestrian injuries [11].

Since the publication of Google's landmark paper "*Attention is all you need*" in 2017 [12] and the subsequent rise of Large Language Models (LLMs) [13], the scope in which AI is being deployed within safety-critical systems has grown significantly. The rapid scaling of capabilities, driven by increasing data and computational resources, shows no signs of slowing down [14][15]. Further, the release of R1 in January 2025, an LLM developed by DeepSeek, suggests that computational techniques beyond scaling pre-training data could lead to more efficient results [16]. Interestingly, work is also well underway to incorporate LLMs as well as multimodal

Authors' addresses: Ibrahim Habli, Centre for Assuring Autonomy, University of York, York, United Kingdom, Ibrahim.Habli@york.ac.uk; Richard Hawkins, Centre for Assuring Autonomy, University of York, York, United Kingdom; Colin Paterson, Centre for Assuring Autonomy, University of York, York, United Kingdom; Philippa Ryan, Centre for Assuring Autonomy, University of York, York, United Kingdom; Yan Jia, Centre for Assuring Autonomy, University of York, York, United Kingdom; Mark Sujan, Centre for Assuring Autonomy, University of York, York, United Kingdom; John McDermid, Centre for Assuring Autonomy, University of York, York, United Kingdom.

variants, such as vision language models, in self-driving vehicles [17] [18]. Vision Language Models (VLMs) use the same transformer-based approach as LLMs but with images rather than text as inputs. Indeed, for safety-critical applications, the tendency of LLMs to produce factually incorrect but seemingly plausible outputs, i.e. AI hallucinations [19][20], presents an open challenge for their use in such applications.

The scale and significance of AI-induced harms make proactive and systematic assurance of safety a priority [21][22][23]. This is particularly the case as AI systems are granted greater autonomy [24][25] and the use of powerful, general-purpose foundational models, or frontier AI, increases [26] [15] [27]. For instance, critical tasks such as clinical diagnoses [28] or driving in open environments [29] are gradually being delegated to AI-based systems. Up until now, the partial automation of such activities has assumed, and is often built around, human oversight as a central risk mitigation measure [30]. The transition of such tasks from human to AI fundamentally challenges long-established engineering standards and safety practices, which centre on high degrees of control of systems and their environments [31][32][33]. It is worth noting that such challenges are also present with the widespread introduction of AI even when human oversight is assumed, due to both technical reasons (e.g. lack of transparency in AI decision making) as well as socio-technical reasons (e.g. an organisational culture where people lack psychological safety to disagree with AI decision making, which hinders their ability to exercise oversight) [34].

For six decades, safety cases have been an accepted means for assuring the development, deployment, maintenance and decommissioning of safety-critical systems across many sectors, most notably nuclear, defence and automotive [35][36][37]. Safety cases provide a way to communicate a clear, comprehensive and defensible argument that a system is acceptably safe to operate in a given context [38]. They consist of a structured argument, supported by a rigorous body of evidence. Importantly, the use of safety cases represents a shared mindset and understanding of how safety should be managed and evidenced, extending beyond merely viewing the safety case as a document [39]. There is growing interest in appraising the suitability of safety cases for supporting the assurance of AI-based systems and services [40] [41]. There has been substantial research undertaken into the development of compelling safety cases for narrow AI models, especially those used in highly autonomous applications [42]. Further, with the remarkable advancement of frontier AI models, particularly LLMs, there has been growing interest in the development of safety cases for such general-purpose AI systems [43]. This has focused almost exclusively on the consideration and mitigation of large-scale risks associated with the development of unauthorised and unacceptable capabilities of LLMs [44]. This work is significant and essential. However, major gaps remain to integrate this fast emerging work into the wider consideration of immediate and imminent risks of harm to individuals and society caused by the integration of general-purpose AI models into safety-critical products and services.

In particular, as with any technology [45], it is essential to consider the risk of harm when using a general-purpose AI as part of a wider system to undertake specific tasks. The safety cases created must take into account the broad operational context to ensure that different kinds of hazards, both immediate and long-term, are addressed in an integrated and consistent manner. In this way, the safety cases are grounded in the established principles of system safety that are expected for all safety-related applications [46].

In this paper, we propose a systematic approach to assuring the safety of AI through the provision of *a whole system safety case.* Specifically, this paper proposes a *Balanced, Integrated and Grounded (BIG) argument* that addresses AI safety at both the technical and the sociotechnical levels. The approach takes an architectural approach to AI safety cases, demonstrating how an overall AI safety argument, in its different social, ethical and engineering aspects, comes together.

The BIG argument primarily builds on the following patterns and methodologies:

- Principles-based Ethics Assurance (PRAISE) [47]
- Assurance of Autonomous Systems in Complex Environments (SACE) [48]

- Assurance of Machine Learning for use in Autonomous Systems (AMLAS) [40]
- Emerging safety argument patterns for frontier AI models (e.g. [44][49])

Overall, the goal of the BIG argument is to improve transparency and accountability for the safety of AI, and ultimately contribute to the development, deployment and maintenance of justifiably safe AI systems. The target audience of this work is wide and diverse:

- Safety specialists and regulators, including those who have to sign-off safety cases;
- AI developers and system/software engineers, covering the full engineering lifecycle;
- Non-technical professionals, including managers, lawyers, ethicists, policy makers and social scientists; and
- Users, varying from trained professionals, e.g. clinicians and pilots, to the general public whose safety is at stake.

The paper is organised as follows. We first provide an overview of safety cases. Next, we present our BIG safety argument, followed by a detailed description of its core sub-arguments (ethical/social, engineering and technical). To illustrate its application, we include examples from diverse domains and AI technologies of different capability, autonomy and criticality. Finally, we conclude with overarching themes for future work.

2 A VERY BRIEF OVERVIEW OF SAFETY CASES

A highly-cited definition of safety cases comes from the UK Defence Standard 00-56, in which a safety case is described "*a structured argument, supported by evidence, intended to justify that a system is acceptably safe for a specific application in a specific operating environment* [38]". The more critical, novel and uncertain the system and its context, the more detailed the safety argument and evidence are expected to be [50].

Safety cases were first adopted in the UK nuclear industry in 1965 in response to the Windscale fire accident in 1957 [51]. The adoption of safety cases was part of a shift in regulation from compliance-based to goal-based approaches. The thinking behind this is that while the regulator can set goals, the developers and operators of safety-critical systems are best placed to determine the means through which these regulatory goals can be achieved, especially in industries that are fast changing and where prescriptive standards would be at risk of lagging behind the state of the art [52][53]. Following major accidents in other industries, the goal-based regulatory approach supported by safety cases was adopted across UK safety-critical industries, including offshore oil and gas production (Piper Alpha explosion in 1988 [54]) and railways (King's Cross escalator fire in 1987 [55] and Clapham main line derailment in 1988 [56]). The construction industry is the most recent domain (high-rise residential buildings) to adopt safety cases following the Grenfell Tower fire in 2017, where 72 people died [57].

When used as a proactive approach to safety management, the use of safety cases has the potential to offer significant benefits [37]. Most importantly, safety cases can provide assurance to developers and operators of safety-critical systems that they have properly understood relevant risks and that these are sufficiently managed [54]. In addition, safety cases as an approach should not be mistaken as simply a written document, i.e. a safety case report, but rather as a structured way and shared understanding of thinking about and managing safety [39]. The mindset underpinning the safety case approach is arguably its greatest strength as it encourages proactive, continual engagement with safety in an open and transparent manner.

However, when safety cases are applied without the accompanying shift in mindset, there is an acknowledged risk that the approach can degenerate into a paper-based bureaucratic exercise that offers little towards improving safety [58]. This was highlighted in the review following the loss of a Royal Air Force Nimrod aircraft in Afghanistan in 2006 [59]. Furthermore, critics of the safety case approach have highlighted that there is a lack of robust evidence about the effectiveness of safety cases as a regulatory approach [60][61][62]. In practice, safety cases are adopted based on face validity rather than conclusive evidence of their utility [37].

For over three decades, research in safety cases, mostly conducted in collaboration with industry, has focused on several key areas:

- Notations: Mostly graphical representations (and also others [63][64]), primarily using the Goal Structuring Notation (GSN) [65] and Claim-Argument-Evidence (CAE) [66];
- Processes: Integrating safety assurance into design from the earliest stages of development and throughout [67];
- Model-based engineering: Providing metamodels and ontologies for model-based safety cases, utilising model transformation, traceability and validation [68];
- Safety case automation: Tool support for argument generation, integration and evaluation [69][70];
- Formal reasoning: Representing safety arguments in mathematical formats to improve precision and consistency [71];
- Modular safety cases: Promoting separation of concerns and enabling compositional representation and analysis [72][73];
- Confidence assessment: Evaluating different kinds of uncertainty, both qualitatively and quantitatively [74][75][76][77][78];
- Reuse: Promoted through argument patterns and templates [79][80];
- Evolution and updates: Addressed via the concept of dynamic safety cases and phased safety case development [81][82][83][84];
- Argument-based assurance: Extended to cover properties beyond safety [85][86], such as security [87][88], ethics [47], and trustworthiness [89].

The current literature already covers some methods and case studies on the use of safety cases for AI systems. Most of these studies often have a narrow focus, such as a tightly-scoped technical problem (e.g. robustness of neural networks for pedestrian detection [90]) or a specific legal consideration (e.g. liability for misdiagnosis using a clinical AI tool [91][92]). Having a narrow focus offers advantages, such as providing stronger and more detailed evaluation evidence. These studies are typically driven by specific requirements in safety standards or guidelines for producing safety cases, most notably in automotive [93][94][95], healthcare [96] and aviation [97][98].

More recently, safety cases have been considered in the context of frontier AI assurance. Unlike existing literature on safety cases that looks at AI systems in specific contexts and for clearly defined purposes (i.e. downstream AI safety [99]), the emerging work on frontier AI safety cases takes a capability-based approach in isolation of a specific deployment environment (i.e. upstream AI safety) [43][41]. The main questions of interest here are: Does a frontier AI model pose a dangerous threat or hazard to society, and is it controlled or controllable? A key focus of these arguments is capability misuse by '*bad actors*', say for the purpose of compromising cybersecurity or producing bioweapons leading to catastrophic or even existential harm.

While the term "capability" is widely used in the literature, a concrete definition is rarely provided. Instead, works commonly present a limited set of examples of capabilities, leaving readers to infer their own definition. Without a clear definition, it becomes difficult to develop approaches to tackle this important problem or determine whether solutions would apply to capabilities more broadly. One definition of capabilities has been supplied in the 2025 International AI Safety Report [100] as "*The range of tasks or functions that an AI system can perform and the proficiency with which it can perform them*". This definition highlights the need for a system-centric approach to AI safety. AI systems are typically orchestrations of frontier models, machine learning components, traditional software, hardware and human operators. As such, assuring the safety of such systems requires more than a consideration of the technological features of AI models. We must also consider the context into which they are to be deployed, as well as the social and ethical landscape. What is needed is a balanced, integrated and grounded argument.

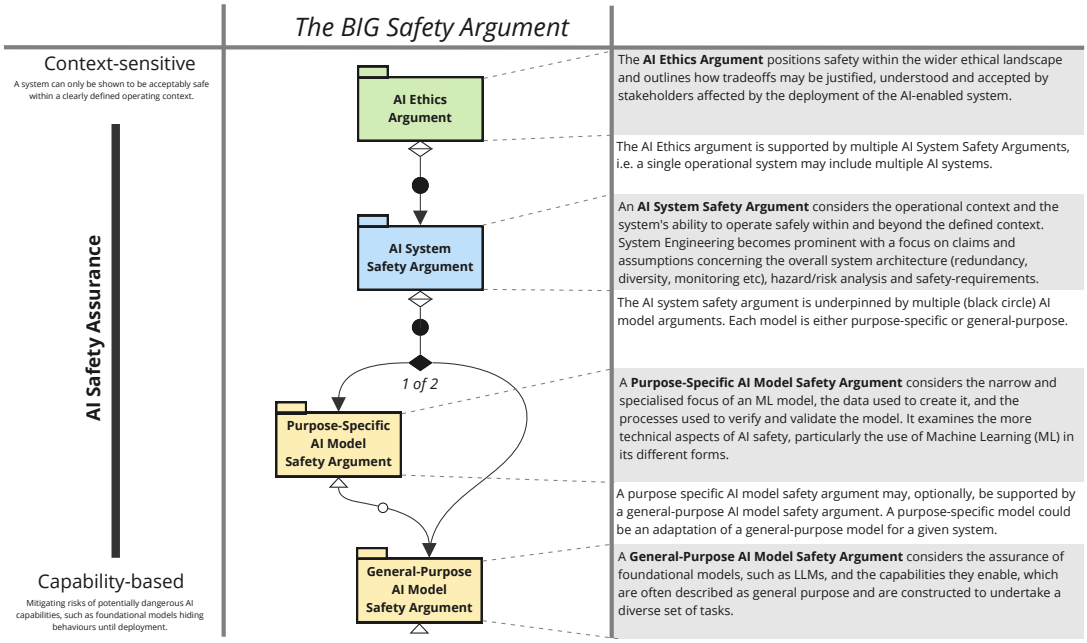

Fig. 1. The Balanced, Integrated and Grounded (BIG) argument, represented in GSN

3 THE BALANCED, INTEGRATED AND GROUNDED (BIG) ARGUMENT

The fragmented state of the literature on AI safety does not help address a central question [15]: how does the whole safety argument come together?

We address this challenge by proposing our Balanced, Integrated and Grounded (BIG) argument for AI safety cases. It is centred on three key characteristics:

(1) **Balanced argument:** Protection from or avoidance of harm is a fundamental ethical principle enshrined in professional codes and legal standards. This principle of safety, or *non-maleficence*, is one part of a set of core principles that should guide critical decision-making about complex societal and technological interventions. Balancing these principles often requires careful consideration and trade-offs [101][47]. For AI systems, a clear and explicit safety argument should facilitate a robust and inclusive dialogue among relevant stakeholders or their representatives, addressing their perspective on safety risks, relating them to other critical issues such as privacy and bias [102][103]. This dialogue should form the basis for a proportionate approach to risk reduction, ensuring that AI safety is not unjustifiably achieved at the expense of other ethical values like fairness [104] and human autonomy [105].

(2) **Integrated argument:** The technical side of AI safety cannot be considered in isolation. Safety assurance must integrate the relevant ethical, social and engineering dimensions of AI deployment [106]. Claims, assumptions and evidence for safety in one dimension can refine those in another. Consider a clinical reinforcement learning system used to support sepsis treatment in intensive care units [96][107]. Claims about mitigating the *clinical risk* of adverse outcomes, such as unnecessarily recommending a vasopressor for a patient with high blood pressure, are intertwined with the *technical objective* of penalising the underlying model for exhibiting such risky behaviour. These clinical and technical issues are interconnected and should be reasoned about in a traceable and integrated manner.

(3) **Grounded argument:** Key aspects of AI assurance will focus on unique issues related to the AI lifecycle and design, such as the completeness of training and testing datasets, model performance and robustness and output explainability [108]. However, to be meaningful in a safety context, these issues must be linked to and grounded in established safety norms [109] such as hazard-oriented safety requirements, risk-driven controls [110] and just safety cultures [111][112]. While AI presents significant and novel safety challenges, these traditional norms and practices remain relevant. In fact, the complex nature of AI systems further emphasises the need for rigour [113] and transparency [114] in how risk acceptability [115][116] and safety requirements are argued about, reviewed and challenged [117].

4 SKETCHING THE BIG ARGUMENT

Figure 1 sketches our BIG AI safety argument. The argument structure, represented using the patterns and modular features of the Goal Structuring Notation (GSN), incorporates the following sub-arguments:

- **AI Ethics Argument:** This argument considers safety amongst other principles that must be assured for the ethical use of AI. Trade-offs between these principles will often be inevitable. The argument outlines how tradeoff decisions and assumptions may be justified, understood and accepted by affected stakeholders or their representatives (e.g. regulators). Our Principles-based Ethics assurance (PRAISE) patterns provide a basis for this argument [47].
- **AI System Safety Argument:** This covers the wider system in which an AI model is deployed, which could be physical (e.g. an autonomous vehicle) or procedural (e.g. drug discovery). The argument details claims and assumptions about relevant hazards and risks, and evidence for the suitability of system-level risk controls including redundancy, diversity, monitoring and meaningful human control/oversight. Our Safety Assurance of Autonomous Systems in Complex Environments (SACE) patterns provide a basis for this argument [48].
- **Purpose-specific AI Model Safety Argument:** This represents claims and evidence for AI models trained and tested for a specific purpose, such as identifying pedestrians using a convolutional neural network or recommending a treatment using reinforcement learning. A key focus of the argument is the justification of the training and testing datasets and the allocated safety requirements and metrics. Our Assurance of Machine Learning for use in Autonomous Systems (AMLAS) patterns provide a basis for this argument [40].
- **General-Purpose AI Model[1] Safety Argument:** This represents safety claims about general-purpose models and capability-specific risks and guardrails, and the supporting evidence, particularly from evaluations (evals), independent audits and red teaming [119][120]. Preliminary argument templates and example safety cases have started to emerge, with focus on specific capabilities, e.g. the ability of a frontier AI to hide its behaviours until deployment or undermine oversight [121][122]. A recent report by the UK AI Security Institute stated that "*Frontier AI safety cases should make arguments about more than just the technical system*" [123]. The BIG argument advances this proposition. It offers traceable means for integrating the assurance of the technical capabilities of these advanced models with a wider set of sociotechnical issues at both the system and societal levels [124][27][125].

[1]Here we use the term General-Purpose AI as defined in [118] i.e. a model created without an explicit consideration of the final system within which the model will be deployed. Such models can typically, without substantial modification, be *tuned* to meet the needs of a specific role within a system.

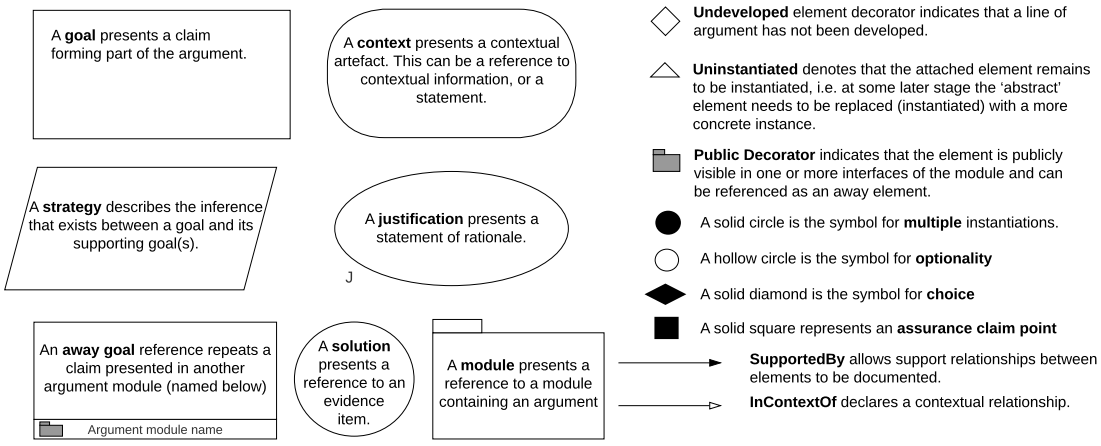

Fig. 2. Symbols and elements of a GSN argument. Extracted and adapted from Assurance Case Working Group [65]

It is important to emphasise that when these General-Purpose AI capabilities are made explicit,[2] they often conflate societal and systems thinking rather than merely considering the technical capabilities of the model. This conflation must be disentangled through a structured argument, as illustrated by the BIG argument, before sufficient action can be taken to address the arising safety issues.

Structurally and notationally, the overall GSN argument pattern is described as follows (see Figure 2 for a GSN legend): The *AI Ethics Argument module* is supported by one or more *AI System Safety Argument modules* (the multiplicity is represented by the solid circle). This indicates that multiple AI systems may require separate safety arguments within a broader safety case. For example, a ground-based system and an airborne system might each need a separate safety argument as part of an overall aviation safety case for an AI-enabled navigation capability. These modules are also 'undeveloped' and 'uninstantiated'. This means that the argument contained within the modules requires further support and specific details (hence its status as a template).

The multiplicity and choice symbols (solid diamonds) under the *AI System Safety Argument module* indicate that this module may be supported by one or more argument modules corresponding to either general-purpose or purpose-specific AI models. As such, the safety argument for an AI model within an AI system is captured in either the *Purpose-specific AI Model Safety Argument module* or the *General-Purpose AI Model Safety Argument module*, depending on its intended use (i.e. general or specific purpose). As purpose-specific AI models often build on capabilities provided by general-purpose AI models, an optional link (represented by the hollow circle) can be used to connect the two corresponding argument modules. It is important to note that Retrieval Augmented Generation (RAG) [127] is increasingly utilised to improve the accuracy of general-purpose AI models for specific purposes [128]. In such cases, the use of RAGs could be justified within the *Purpose-specific AI Model Safety Argument.*

[2]In [126], a capability is suggested as "*Capable of significantly enabling a non-expert to develop known biothreats that could increase their ability to cause severe harm compared to other means*". Here the role of the model within the larger system may be considered as a publicly accessible knowledge source, or natural language search tool. For harm, of this type, to occur requires a societal context in which the non-expert has an adversarial mindset and access to the resources necessary to make use of the knowledge supplied. To mitigate this issue at the technical level requires a consideration of issues such as data poisoning and deployment mitigation strategies in the wider software framework. The General-Purpose AI model does not, on its own, have the capability of producing harm.

Finally, in this paper, we focus on the structural aspects of the BIG argument. However, a safety case for a complex intervention such as AI is rarely static. Rather, it is a living, dynamic approach integrated into the wider design and operational processes. This approach evolves with new evidence and an updated understanding of the system's actual performance in its intended operational environment [81].

The next three subsections explore the argument modules in more detail. Additional practical guidance, and the underpinning assurance methodologies, are detailed in [47][48][40][3].

4.1 AI Safety: The Ethical Argument

We situate the top-level argument within the wider ethical landscape [129]. This is important for three reasons. Firstly, ensuring safety is a fundamental ethical obligation. Secondly, claims about AI safety are inseparable from claims about other ethical principles such as AI fairness and transparency. Thirdly, tradeoffs between various ethical principles (such as transparency and privacy) are often necessary and therefore have to be explicitly considered, justified, challenged and where appropriate accepted.

Here, we build on the four classical principles of biomedical ethics [130]. These are the principles of justice, beneficence (do good), non-maleficence (do not harm) and respect for human autonomy. As we argue in more detail in [47], these principles provide a plausible normative basis and coverage of key ethical values such as sustainability, dignity and reciprocity [131][132][133]. Burr and Leslie present an alternative, bottom-up, approach to structuring an AI ethics argument [89].

The argument for each of the four principles is contained within a separate module in Figure 3. The '*Ethics Assurance Argument*' module captures the overall argument for ethical acceptability, which is centred on making the case for the *just* deployment of the AI system. This is detailed in the '*Justice Argument*' module and appeals to the equitable distribution of benefits and harms across all affected stakeholders. This argument deals with the necessary issue of resolving and justifying tradeoffs and builds on and integrates separate and detailed arguments concerning the benefits offered by the use of the AI system (through the '*Beneficence Argument*' module) as well as the mitigation of harm posed by it (via the '*Non-maleficence Argument*' and '*Human Autonomy Argument*' modules).

It is important to note that the consideration of safety is not limited to the '*Non-maleficence Argument*' but cuts across all argument modules. For example, respecting human autonomy, covered in the '*Human Autonomy Argument*' is fundamental for assuring effective oversight. Otherwise, the role of humans as a risk control in AI-based decision support systems is weakened. This in turn may undermine confidence in the overall safety case. Similarly, reasoning about proportionality, which is central for making decisions about risk acceptability, is considered in the '*Justice Argument*' module, as it often hinges on some form of risk-benefit analysis, including trade-offs.

Another central aspect of the ethical AI debate is transparency [134], captured in the '*Transparency Argument*' module, which is essential for safety assurance. In this argument, we consider transparency as a supporting principle. It plays two key roles in the overall AI safety argument.

- Firstly, it presents transparency claims about the AI development, supply-chain and deployment processes, e.g. why the training datasets were selected and how they were preprocessed to ensure accuracy and balance.
- Secondly, it directly links to explainability of the AI outputs and the extent to which the specific formats and modes of explanation are appropriate and meaningful for the intended recipients (e.g. feature importance vs counterfactual reasoning)[135]. We specifically build on the philosopher Paul Grice's maxims of cooperative communication [136][137]. The four maxims of quantity, quality, relevance and manner

[3]Some aspects of PRASE, SACE and AMLAS have been adapted and abstracted to ensure consistency and brevity. Readers are advised to consult the primary sources [47][48][40] for a full description of the methodologies and argument patterns.

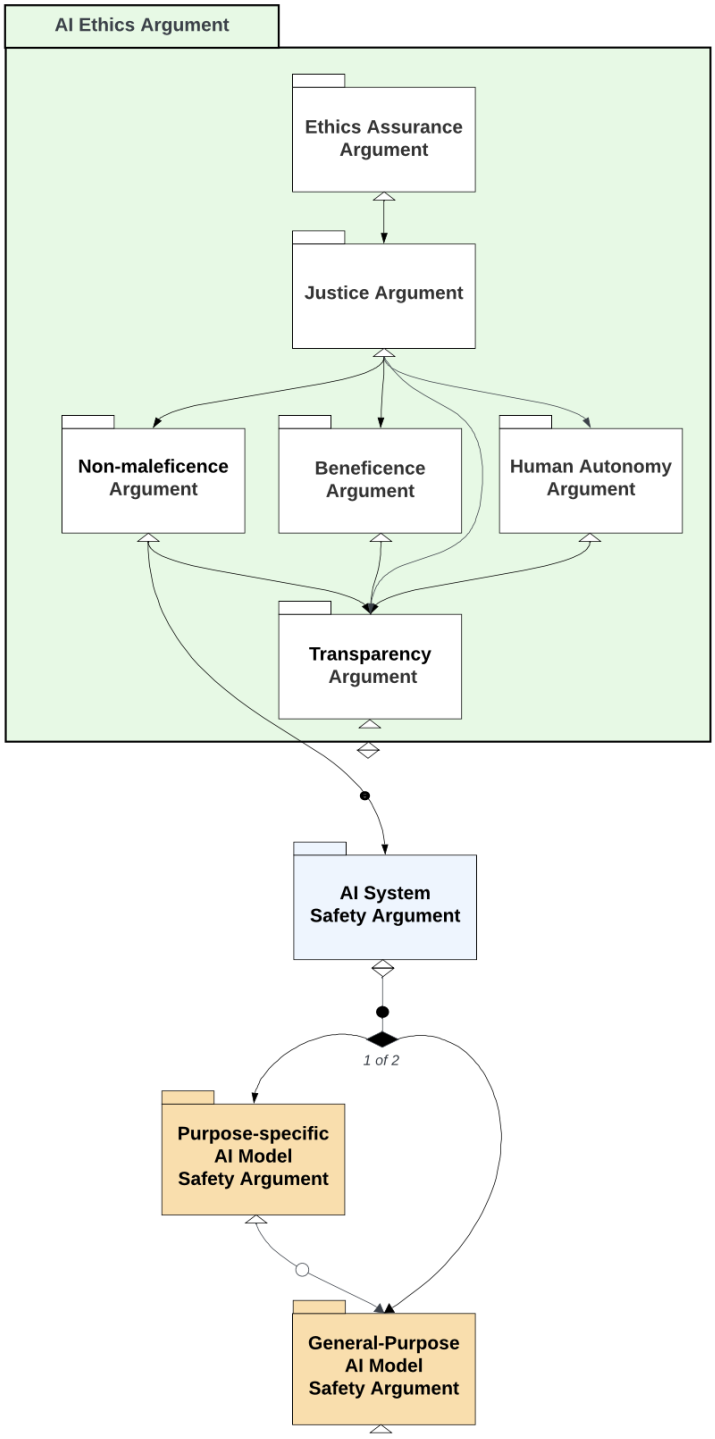

Fig. 3. The Ethical Argument represented in GSN

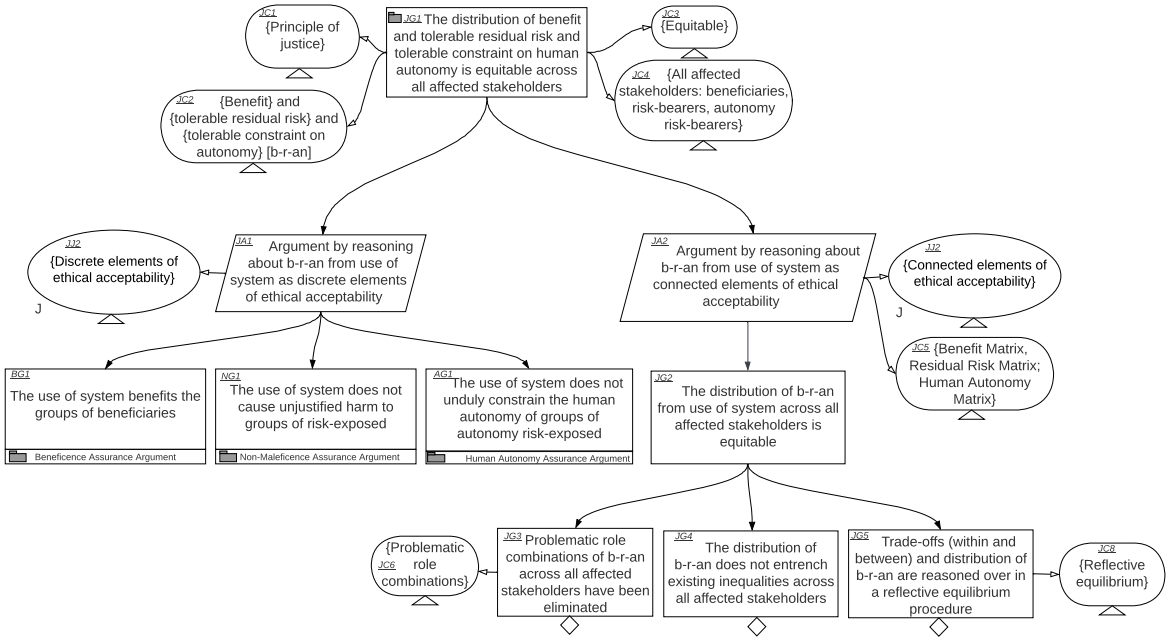

Fig. 4. Justice argument module of the PRAISE pattern represented in GSN [47]

provide a sound basis for assuring the suitability and effectiveness of communication about critical AI properties.

The *AI Ethics Argument* is described in full in [47]. However, to illustrate how this argument brings the different ethical considerations together, we next discuss the *Justice Argument* in more detail. The argument contained within this module is shown in Figure 4. The top claim states that the "*distribution of benefit, tolerable residual risk, and tolerable constraint on human autonomy is equitable across all affected stakeholders*".

This claim appeals to the notion of distributive justice, reflecting a deeper consideration of who benefits and who bears the risks from the use of AI. This in turn provides a more transparent basis for judging the proportionality of risks to different stakeholders and at what cost or benefit to others and themselves [47].

To support this claim, the argument is built on two overarching strategies. The first, *JA1*, details claims that the system provides benefits, does not cause unjustified harm and does not unduly place constraints on human autonomy. This strategy considers these issues as discrete elements of ethical acceptability, where each is developed further in a separate argument module. The second strategy, *JA2*, collectively considers benefits, risks of harm and constraints on human autonomy, focusing on inevitable tradeoffs and their justification (JG5), e.g. through reflective equilibrium [138], associated with John Rawls' work on justice [139]. Two additional claims are emphasised in this argument: that unacceptable role combinations are eliminated (*JG3*) e.g. certain groups only bearing risk, and no benefit, from AI's use, and that AI deployment does not entrench existing inequalities (*JG4*).

A key challenge in supporting the *Ethical Argument* lies in the limited availability of practical methods and techniques for translating ethical principles into concrete requirements [129]. Example 1 illustrates a potential approach to addressing this challenge.

Example 1. Identifying Ethical Concerns for an AI-enabled Assisted Dressing Robot

AI-enabled robotics have been proposed as a way of improving the lives of those living long-term conditions which restrict physical capabilities. Deploying such systems could increase independence in the elderly and reduce the need for traditional care services. The services offered by such systems will need to be personalised to the user and the context in which they are to be deployed.

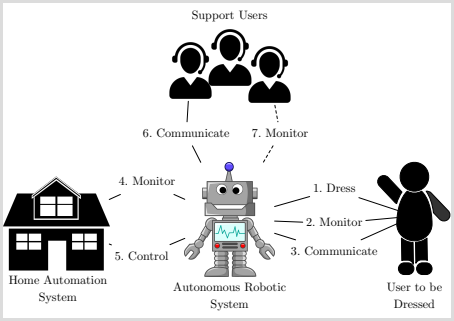

Fig. 5. Robotic assisted dressing platform [140]

Figure 5 shows one such case where an AI-enabled robotic system is deployed to aid a user to dress in their own home [140][141]. In order to undertake the primary function of dressing, the platform will need to monitor the user and communicate with them. Since dressing necessitates the user to be undressed (or partially dressed) the platform is also able to control the heating, curtains and lighting in the home. Finally, since the user may be vulnerable, facilities are available to communicate with a support team who, in turn, can monitor the state of the system to ensure it is functioning as expected.

For critical systems such as this, it is necessary for us to derive requirements which are not only functional in nature, but also respect Social, Legal, Ethical, Empathetic and Cultural (SLEEC) norms [140]. These norms are derived from high-level principles (Table 1) and refined through a structured elicitation process (Figure 6) to define rules and address the trade-offs arising from the context into which the system is to be deployed and the multi-disciplinary requirements of system stakeholders.

SLEEC Concern	Description
Privacy	Limiting intrusion on the personal space of the user and ensuring privacy is protected; safeguarding health data, practising good data stewardship, and granting or restricting access to medical records
Respect for Autonomy	Granting and withdrawing of permissions, including consent and assent; ensuring the user maintains an appropriate level of control
Dignity	Understanding and accommodating the user's social and cultural sensitivities, respectful treatment
Explainability and transparency	Informing the user about system decision-making and any inferences made; providing justification for a course of action adopted
Beneficence	Maximising good outcomes
Non-maleficence	Minimising harm by ensuring safety and reducing the possibility of physical and psychological harm to the user

Table 1. Examples of SLEEC concerns in the robotic asssisted dressing system [140]

The complexity of the rules arising from such systems could lead to conflicts and redundancy. The SLEEC methodology may support the *AI Ethics Argument* by providing a potential approach to identifying and analysing key ethical issues and how they relate to other social norms. It offers a language, formal semantics and a toolset to encode these rules for use in robotic and AI systems and part of the evidence base needed to support the top argument.

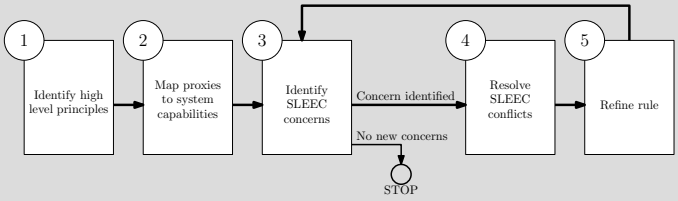

Fig. 6. The SLEEC requirements elicitation process [140]

5 AI SAFETY: THE SYSTEM ARGUMENT

This argument refines the ethical claims about AI safety, considering them in more detail within both the context of the wider system (e.g. a maritime autonomous surface ship [142][143]) and the social and organisational setting in which it is deployed (e.g. passenger ferry services in busy maritime environments and staffing levels that impact on the ability of people to augment autonomous AI functions).

Issues of particular focus are the scope of the operational domain, considering the system's ability to safely operate within and beyond its defined context, and human-machine interactions, including challenges around over- or under-reliance on AI-enabled functions. The systems engineering perspective becomes more prominent here, focusing on claims and assumptions about the overall system architecture, including redundancy, diversity and monitoring (see Example 2: architectural tactics for incorporating AI in aviation). It also addresses how detailed safety requirements, incorporating uncertainties, are determined, refined, and verified.

Example 2. Architecting AI-based Autonomous Aviation Systems

When using AI in safety-critical systems it is crucial to take a systems perspective to understand and manage its contribution to risk. Architectural patterns can be used to mitigate some AI failure modes or uncertain behaviours. For example, it may be possible to monitor and constrain outputs from an AI-based drone flight stabilisation component to prevent unexpectedly large directional changes being sent to propellers. Such outputs could physically strain the drone and/or lead to sudden and unpredictable trajectories around infrastructure [144]. A monitor-based architecture thus reduces the AI component's individual contribution to risk of collision and improves reliability of the overall system. Further, it may make assurance requirements on the AI model safety argument less onerous.

There are many different architectural design patterns (e.g. component monitoring, component switches, voting on outputs from multiple diverse components with the same function) which can be combined to help incorporate AI into avionics systems and maintain existing high-assurance norms [98]. For example, runtime monitors can capture information on real-time performance of AI components, switching to a traditional (but perhaps less adaptive) alternative function if the performance is below a particular threshold (Figure 7)[145]. These patterns of architectural designs have been used for avionics systems for many years, but AI provides additional challenges to their efficacy. For example, run-time performance of an image classifier is typically difficult to accurately assess due to the lack of ground truth for comparison.

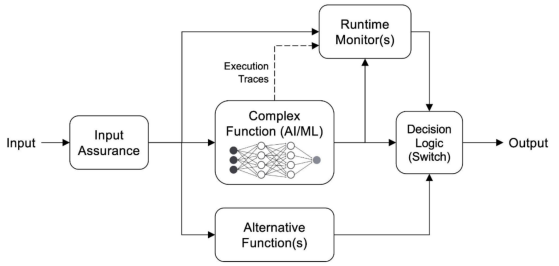

Fig. 7. Runtime assurance (RTA) architectural pattern [98]

Figure 8 depicts the '*AI System Safety Argument*' module. The top-level claim (*G0*) is that the "*{AI System (AIS)} is sufficiently safe throughout its entire operational life*", with the curly brackets indicating that the term '*AI System (AIS)*' requires instantiation. The argument strategy supporting this claim centres on the ability of the system to remain sufficiently safe within (*G1*) and outside (*G7*) its defined operating context (see Example 3: modelling intravenous infusion administration in intensive care units). Following that, the argument takes a hazard-based approach (*G3*), focusing on mitigating the identified hazardous scenarios by developing safety requirements and constraints on the operation of the system. Collectively, these requirements and constraints constitute the Safe

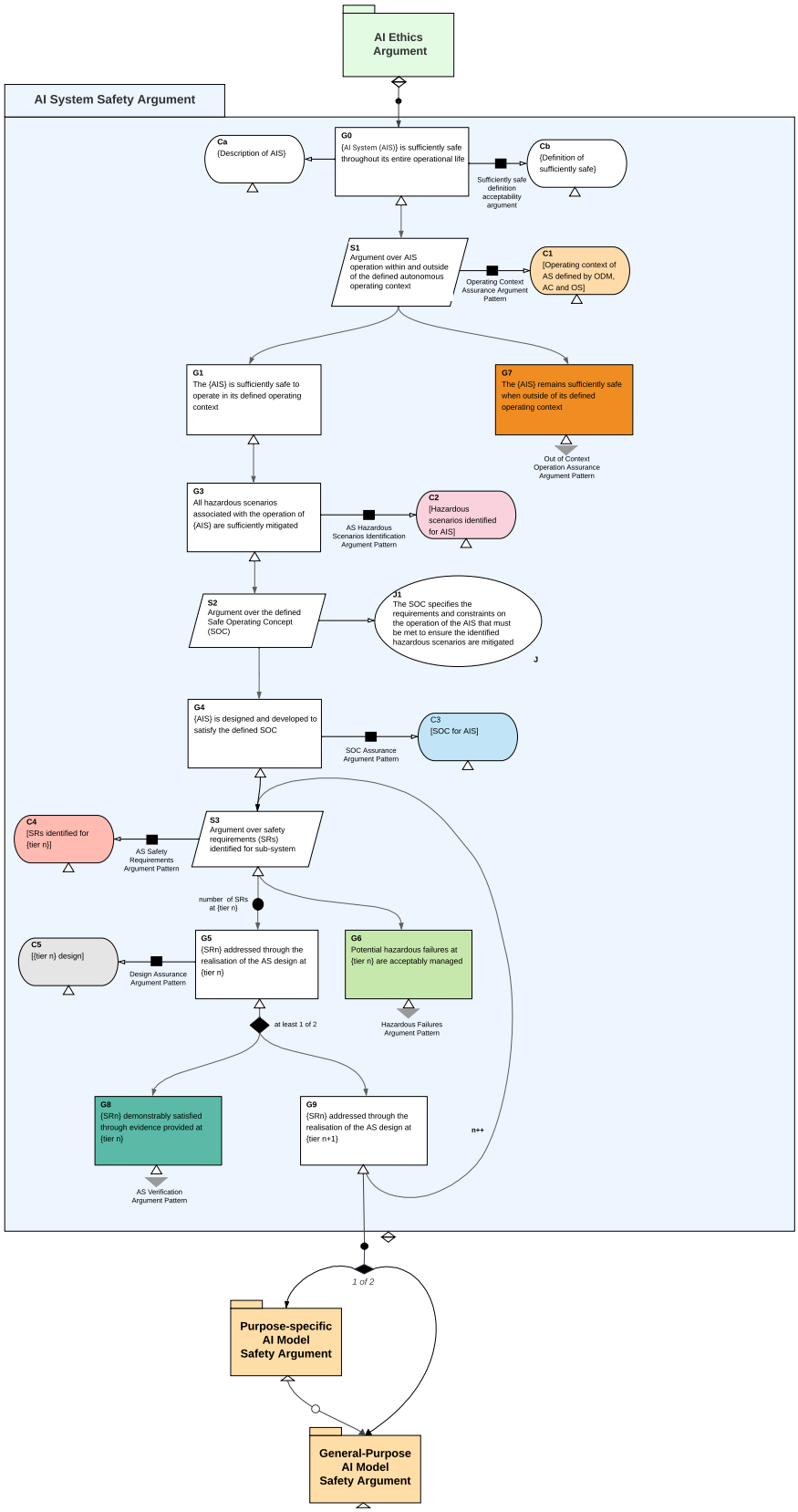

Fig. 8. The AI System Argument represented in GSN

Operating Concept (SOC) (*J1*). The argument is iterative in nature, decomposing assurance claims about the development and refinement of the safety requirements at different levels of abstraction, as indicated in *G9* and *S3* (see Example 4: generating AI safety requirements for satellite-based wildfire detection).

Example 3. Complex Contexts: Performance Variability in Intravenous Infusion Administration in ICU

The use of AI in healthcare is high on the political agenda. However, to meaningfully incorporate AI-based tasks into clinical pathways, it is important to model and analyse existing clinical needs, challenges and constraints. Appreciating the complexity of clinical practice and reducing the gap between work-as-done and work-as-imagined is essential [146]. FRAM (Functional Resonance Analysis Method) is a well-established technique in safety science used to model the performance variability of complex sociotechnical systems (work-as-done) [147].

Figure 9 shows a FRAM diagram of intravenous infusion administration [148]. This links to the need to define the operating context of the system (Context *C1* in the *AI System Argument*). The purpose behind the model and analysis, detailed here [148], was to ensure sufficient understanding of current practice as a prerequisite for automating any tasks using AI systems.

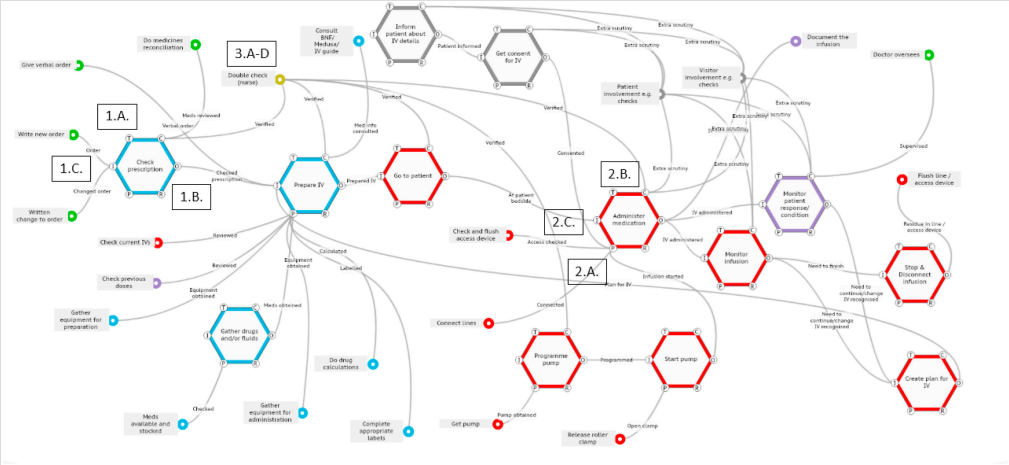

Fig. 9. FRAM network diagram of intravenous infusion administration [148]

For example, the model indicates that some variability in medication ordering in ICU may be justified. The standard operating procedure (SOP) suggests nurses should always have a written prescription beforehand to ensure correct administration. However, emergencies may require immediate drug administration, or doctors may be too busy to write an order immediately and advise proceeding without it, on the understanding that the prescription will be issued later. This illustrates how people make adjustments in everyday work in order to deliver care successfully given competing demands and priorities. When designing and deploying AI-based systems, it is important to consider whether and how their use might potentially disrupt people's ability to make such adjustments flexibly, e.g. if the AI requires that a prescription has been issued on the electronic system without the ability to afford any flexibility.

It is important to note that Assurance Claim Points (ACPs), represented as black squares, are used in the representation of this argument [75]. ACPs provide links to confidence arguments that justify the sufficiency of confidence in specific aspects of the safety argument. For example, for a hazard-based argument such as the one presented in Figure 8, confidence that all hazardous scenarios associated with the operation of the system are identified is fundamental. To this end, an ACP is added to the contextual link between *G3* and *C2*, creating a pointer to a detailed confidence argument concerning the way in which scenarios were generated, reviewed and updated. The full argument is explained in detail in [48].

Example 4. AI Safety Requirements for Satellite-Based Wildfire Detection and Alert System

A satellite with a multi-spectral imager passes over a region of interest that may contain wildfires (Figure 10). An Artificial Neural Network (ANN) onboard the satellite is trained to detect wildfires in the images received and to send an alert to a ground station identifying the location of the fire [149]. This alert can then be passed to the relevant authorities who can respond appropriately. Detecting wildfires onboard a satellite reduces bottlenecks and delays associated with sending image data to be processed on the ground [150].

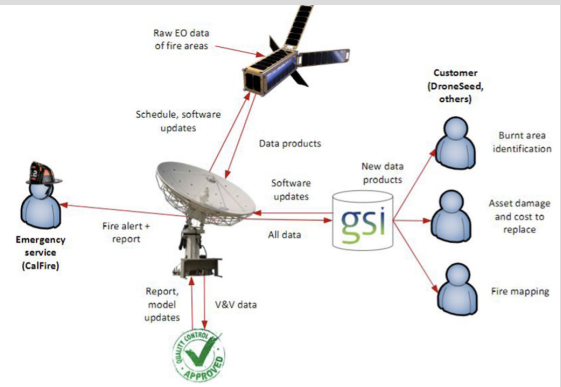

Fig. 10. Concept of Operations for Wildfire Alert System [149]

Two potential hazards were identified for the wildfire alert system:

- **Hazard 1** is that the emergency services miss a wildfire emergency, which could lead to a delay in the response to the fire, a larger and less controlled fire, and thus a potential increase in the risk of harm to people and property or putting firefighting teams in danger. It is determined that the AI-based wildfire alert system could contribute to this hazard through failure to detect the presence of a wildfire. Table 2 shows three safety requirements identified for the wildfire alert system in order to mitigate the contribution of the system to this hazard through the specification of required accuracy and response time of the AI-based wildfire alert system.
- **Hazard 2** is that an alert is raised for a wildfire that does not actually exist. This could result in fire response resources being mis-assigned and thus unavailable to respond to real wildfires in a timely manner. The AI-based wildfire alert system could contribute to this hazard through detecting a wildfire in the incorrect location. Figure 2 shows an associated safety requirement for the wildfire alert system that specifies an acceptable rate of false detections through comparison to current systems.

Table 2. System safety requirements for wildfire alert system [149]

Hazard 1 - Services Miss an Emergency	
REQ-SAFE-ER-1	The Emergency Response Service shall determine the location of an active wildfire within 200m of its true location.
REQ-SAFE-ER-2	The Emergency Response Service shall inform emergency services of an active wildfire within 3 hours of it starting.
REQ-SAFE-ER-3	The Emergency Response Service shall positively identify 95% of all active wildfires acquired by the satellite instrument within the area of interest.
Hazard 2 - Services are Directed to a False Emergency	
REQ-SAFE-ER-4	The Emergency Response Service shall falsely indicate active wildfires in the area of interest at a rate not exceeding current fire alert service (average for FIRMS of 52 per month) .

This safety requirement specification provides Context *C4* to the argument in Figure 8, and justification for the sufficiency of these safety requirements is provided as a confidence argument.

Table 3. Assurance methods for the Data Management stage adapted from [108]

	Associated activities†				Supported desiderata‡			
Method	Collection	Preprocess.	Augment.	Analysis	Relevant	Complete	Balanced	Accurate
Use trusted data sources, with data-transit integrity guarantees	✔				★			
Experimental design	✔		✔		★	★	☆	
Simulation verification and validation			✔		★	☆	☆	
Exploratory data analysis				✔		★	★	
Use adversarial examples			✔		☆	★		
Include a "dustbin" class			✔		☆	★		
Remove unwanted bias		✔	✔		★		☆	
Compare sampling density			✓	✔		★	☆	
Identify empty and single-class regions			✓	✔		★	☆	
Use situation coverage				✔		★		
Examine system failure cases				✔		★		
Oversampling & undersampling				✔		★	★	
Check for within-class and feature imbalance				✔		★		
Use a GAN			✔			★	☆	
Augment data to account for sensor errors	✓		✔		☆			★
Confirm correct software behaviour	✓	✔	✔	✓	☆	★	☆	☆
Use documented processes	✔	✔	✔	✔	☆			★
Apply configuration management	✔	✔	✔	✔	☆			★

†✔ = activity that the method is typically used in; ✓ = activity that may use the method
‡★ = desideratum supported by the method; ☆ = desideratum partly supported by the method

6 AI SAFETY: THE PURPOSE-SPECIFIC AI MODEL SAFETY ARGUMENT

This argument starts to cover the technical aspects of AI safety. In particular, we focus on Machine Learning (ML) in its different forms, such as supervised, unsupervised, and reinforcement learning. The argument considers ML-based functions deployed to serve a specific and often narrow purpose, e.g. diagnosis of particular clinical diseases in specific pathways. The safety claims, assumptions and evidence in this argument cover the entire ML lifecycle, including data curation, model training and testing, and subsequent deployment, monitoring and updates (following the AMLAS methodology [40]). The argument is largely requirements-driven, justifying how system-level safety requirements, considered in the *AI System Argument* above, are broken down into specific technical AI safety requirements. Evidence that these requirements have been validated and verified is then generated within the wider system and environment. Key claims centre on specific performance and robustness metrics, quantified safety thresholds, accuracy and representativeness of the datasets and explainability of the model outputs. Fundamental concerns include *traceability* between the technical claims at this level and the higher-level safety claims at the system and ethical/societal levels.

Here, we build on the AMLAS methodology to structure the pattern for a purpose-specific AI model safety argument. Figure 11 depicts a simplified and adapted composition of the 6 sub-arguments comprising the AMLAS safety argument pattern. These correspond to safety assurance within the following interrelated stages in the ML lifecycle [40]:

(1) ML Safety Assurance Scoping
(2) ML Safety Requirements Assurance

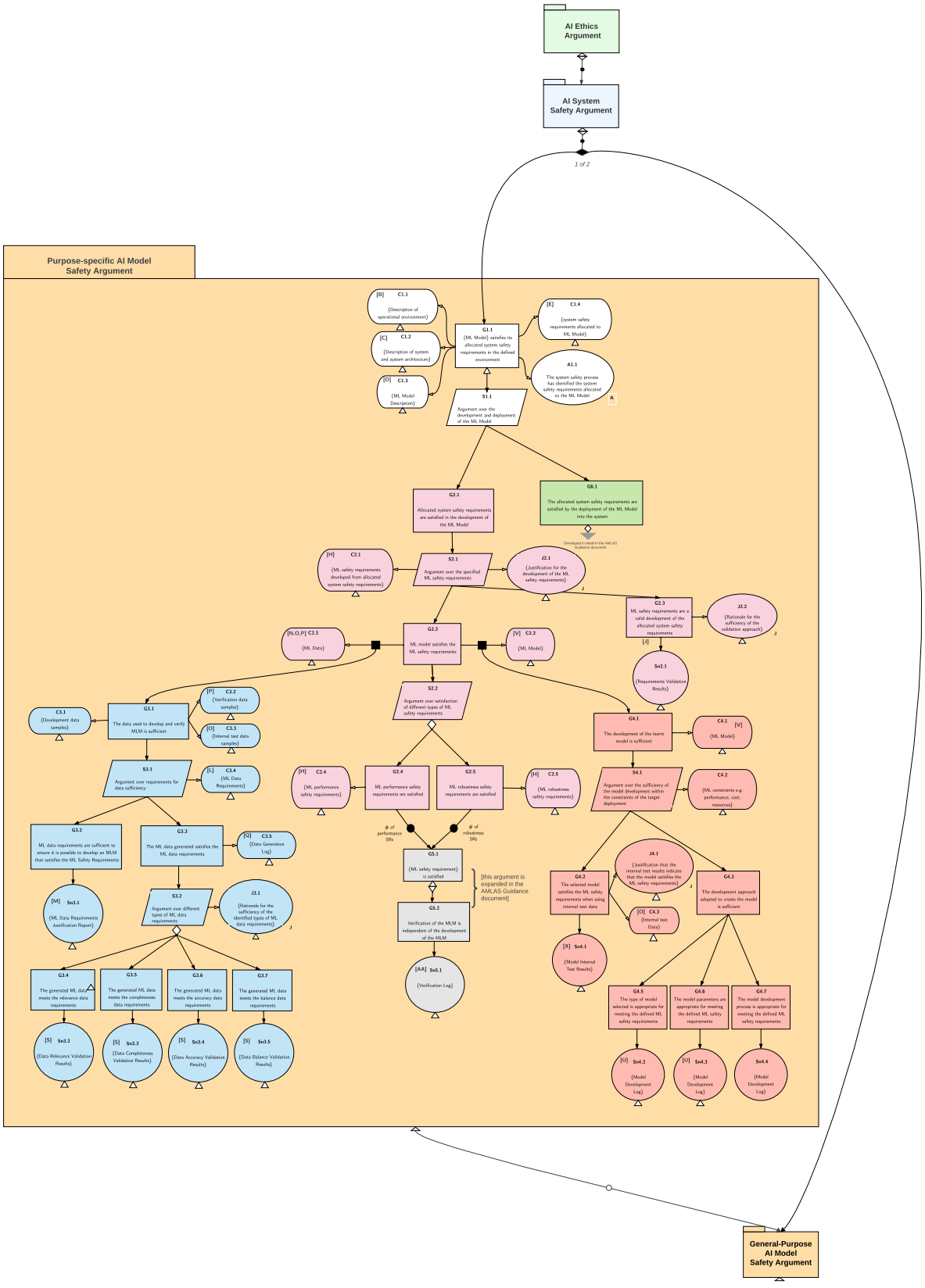

Fig. 11. The Purpose-specific AI Model Safety Argument represented in GSN (bird's-eye view)

(3) Data Management Assurance
(4) Model Learning Assurance
(5) Model Verification Assurance
(6) Model Deployment Assurance

Example 5. Safety-driven Design of Machine Learning for Sepsis Treatment

Sepsis, a life-threatening organ dysfunction caused by a dysregulated host response to infection, stands as one of the leading causes of mortality and one of the most resource-intensive conditions to treat in hospitals. The use of Reinforcement Learning (RL) can help to discover an optimal treatment strategy, particularly optimising the administration of vasopressors and fluids, which are two fundamental medications for sepsis treatment.

Whilst learning the optimal treatment, the RL system also must not learn hazardous behaviours. One of the hazardous scenarios is a sudden change of vasopressor dose, which can cause significant harm to the patients, e.g. resulting in acute hypotension (arising from rapidly decreasing doses), hypertension or cardiac arrhythmias (arising from rapidly increasing doses) [96]. Therefore, we evaluated whether such behaviours exhibited in the original learnt policy, showing that 35% of the cases that the RL model would recommend sudden change in vasopressor dose compared to 3% in clinician policy, i.e. what clinicians have done for the same patient cases, as shown in Table 4.

Table 4. Summary of max dose change between consecutive doses for the three policies

	Dose of vasopressor (mcg/kg/min)	
	Small-Medium Dose Change (0-0.75)	Large Dose Change (>0.75)
Clinician Policy	97% (2,100)	3% (60)
Original Policy	65% (1,404)	35% (756)
Modified Policy	92% (1,990)	8% (170)

Guided by AMLAS, especially the model learning stage, we modified the loss function and the feature space in the RL model, as shown in Table 5, then retrained the RL model. The resulting modified policy showed only 8% of sudden vasopressor dose change when evaluated on the same patient cases, which is much closer to the clinicians' behaviour. This shows the value of following a systematic approach for proactively mitigating the system level risk in the context of the machine learning lifecycle.

Table 5. Major changes in the modified RL model

	Features in state space (R1)	**Cost Function(R3)**
Original RL model	48	$L(\theta) = E[(Q_{double-target} - Q(s,a;\theta))^2] + \lambda_1 max(\lvert Q(s,a;\theta)\rvert - Q_{thresh}, 0)$
Modified RL model	48 (Removed one feature – timestep, added an extra one – relative dose change)	$L(\theta) = E[(Q_{double-target} - Q(s,a;\theta))^2] + \lambda_1 max(\lvert Q(s,a;\theta)\rvert - Q_{thresh}, 0) + \lambda_2 max(\lvert V_{change}\rvert - 0.75, 0)$ V_{change} is the agent recommended dose (argmax of $Q(s,a;\theta)$) minus the vasopressor dose in the previous step; λ_1 and λ_2 are the tuning parameters that decide how much to penalise the flexibility of the model.

The focus on technical development considers ML developers, systems engineers and subject matter experts (e.g. clinicians or pilots) as the primary stakeholders in generating the evidence necessary at this stage. The process outlined in the AMLAS methodology connects the SACE and AMLAS patterns through the safety assurance

scoping argument, which maps the role of the ML model under consideration to the potential for hazardous scenarios that may arise from the functions undertaken by the ML model. As such, the top-level claim (*G3.1*) made by the application of AMLAS is that the "*ML Model satisfies its allocated system safety requirements in the defined environment*". See Example 5, illustrating a traceable link between clinical hazard and the training requirements for an RL agent involved in supporting the treatment of sepsis.

Further, given the data-intensive nature of ML, the argument pattern pays particular attention to the justification of the choice of the training and verification datasets. For instance, the safety claim (*G3.2*) creates an assurance link between the ML safety requirements (traced to system-level safety requirements) and the data requirements. That is, the data requirements are sufficient for realising the system-level safety considerations at the data level. This is then refined to consider the rationale for specific data desiderata, primarily relevance, completeness, accuracy, and balance (see Example 6, illustrating an approach to mitigating the impact of rare subclasses in deep neural network classifiers).

Example 6. Detection and Mitigation of Rare Subclasses in Deep Neural Network Classifiers

Legal frameworks make explicit lists of protected characteristics which allow for us to define measures of fairness against which we can evaluate our systems. These characteristics can be used in the development of data collection activities as well as evaluation processes, e.g. ensuring that gender bias is avoided.

Unfortunately these lists are often too coarse to identify pockets of intersectional data where individuals may still be unfairly treated in practice. Monitoring and mitigating such subclass discrimination requires us to identify rarity within the data used to train our models and observed at run-time [151].

Figure 12 illustrates a process in which data samples are evaluated with a simple commonality score that is correlated to the probability of misclassification, and hence unfair bias, in the resulting AI-enabled system. Samples which are dissimilar to the greater population are then examined to identify common features and mitigations (such as data augmentation of collection) is undertaken to correct for the bias.

A similar assessment of the commonality score at run-time allows us to present this to the user, along with our prediction to allow for adjustments to be made at run time. This method represents an example approach for justifying the quality and coverage of the AI data management process.

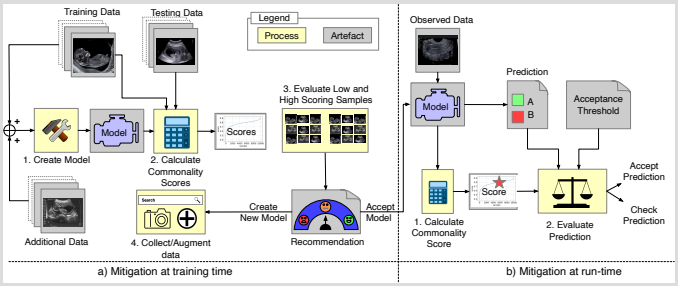

Fig. 12. Processes for mitigating rare subclasses at training time (left) and run-time (right). Adapted from [151]

No single tool or technique can provide comprehensive safety evidence across all stages of the ML lifecycle. However, a systematic approach can use sets of diverse and complementary tools and techniques to create a compelling safety argument for each stage. Table 3 shows how considering tasks involved with data management and the desirable features of data can build evidence [108]. Such evidence supports the claim that the data used to develop and verify an ML component is sufficient (Example 7 provides an example technique for generating such verification evidence).

Example 7. Verification of Contextually Relevant Robustness for Neural Network Image Classifiers

Traditional measures of performance on models used in AI-enabled autonomous systems can provide a false sense of confidence for systems operating in open-world contexts. Table 6 shows a set of nominal accuracy figures for a set of neural networks trained on the CIFAR-10 identification problem. However, these figures alone tell us little about the robustness of the models when deployed and therefore additional evidence that they are suitable for use in these complex contexts is required.

Table 6. CIFAR-10 model accuracy [152]

	Model	Accuracy		Model	Accuracy		Model	Accuracy
4A	Small Relu	49.11	5A	Large Relu	53.20	6A	CNN	84.07
4B		47.45	5B		53.04	6B		85.17

Verification is an important step in providing such evidence. Since vision-based systems have been shown to be susceptible to small perturbations in the input space [153], such an approach may be considered a measure of robustness. However, it lacks semantic meaning. What we need are tools and techniques which are of practical value to ML engineers, allowing them to build mitigations and operational safeguards which are aware of the limitations in robustness of the models used. Figure 13 shows a process by which we may verify contextually meaningful measures of robustness [152].

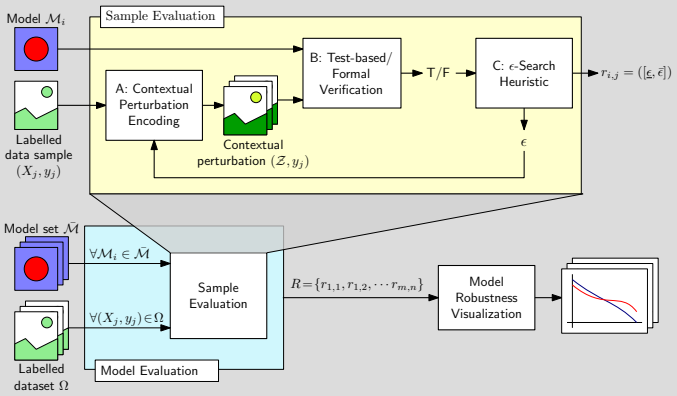

Fig. 13. Process for verifying contextually meaningful DNN robustness [152]

In this work a set of contextually meaningful perturbations are identified, through discussions with domain experts, and a formal encoding of the perturbation defined such that $\epsilon \in [0, 1]$. Data samples are then perturbed with values of ϵ to gradually degrade the samples. We can then identify the level of perturbation necessary for a model to fail for a sample.

Figure 14 shows the result of this process. As the levels of contextually meaningful perturbation (haze, contrast and blur) present in the image increase we see a corresponding degradation in the model performance. We note however that the rate of degradation is not constant across all models, indeed for haze, the 'best model' changes as the image degrades. Such evidence may lead us to refine our model in the presence of such conditions or to deploy multiple models with appropriate monitoring and switching to mitigate safety concerns.

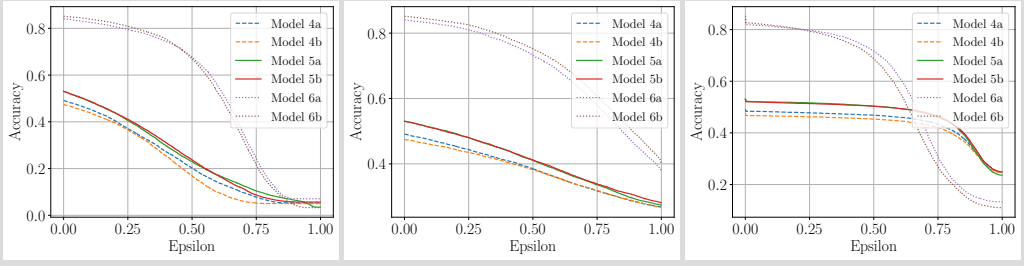

Fig. 14. Plot of model robustness for a set of CIFAR-10 models [152]

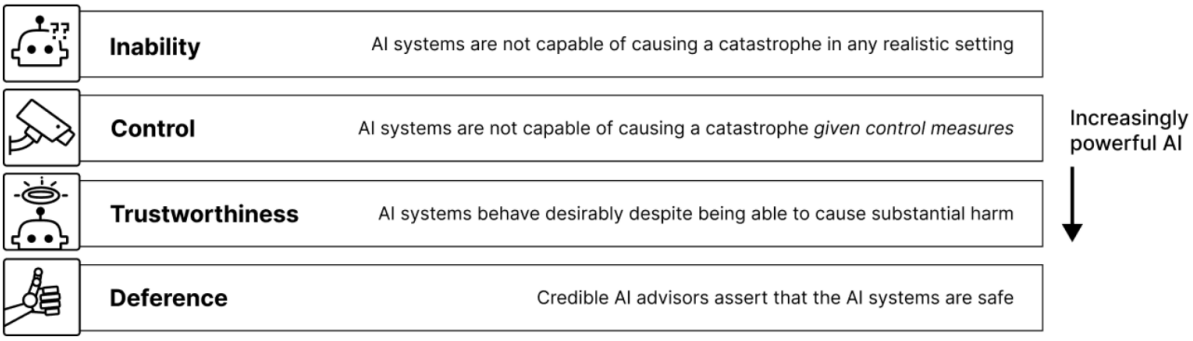

Fig. 15. Building block arguments for making frontier AI safety cases [43]

7 AI SAFETY: THE GENERAL-PURPOSE AI MODEL SAFETY ARGUMENT

At this level of abstraction, the BIG argument considers the rapid adoption of foundational or frontier AI models, such as LLMs, including in critical applications such as healthcare [154][155]. These models are often presented as general purpose, with the intended context rarely specified [15]. This makes their safety assurance at odds with long-established safety principles and practices that consider safety as a context-sensitive property [156] [99].

The general-purpose nature of foundational models has shifted the technical AI safety debate from context to capability [157][158][159]. Major initiatives for assuring the safety of General-Purpose AI (GPAI) models have concentrated on the potential '*harmful*' outcomes that model capabilities may cause. For example, Google DeepMind's Frontier Safety Framework focuses on different critical capability levels (CCLs) [160]. CCLs describe "*protocols for the detection of capability levels at which models may pose severe risks*" [160].

The latest version (2.0) focuses almost exclusively on misuse and deceptive alignment risks. For example, for the former risk, the framework pays a particular attention to frontier models "*assisting in the development, preparation, and/or execution of a chemical, biological, radiological, or nuclear ("CBRN") attack*" [160]. It considers the use of safety cases for assuring the sufficient mitigation of this risk at the development, pre-deployment and post-deployment stages.

The risk associated with these frontier model capabilities may be viewed as a common cause failure or a particular risk [99]. That is, these failure conditions are problematic regardless of the specific context or application (i.e. downstream safety). This has led to an emphasis in the technical AI safety literature on evaluation, independent audits and red-team testing conducted at the model development stage and at the capability level (i.e. upstream safety), where it is believed to be more effective to identify early warning signs (see Example 8 for an independent pre-deployment evaluation of Anthropic's Claude 3.5 Sonnet [161]).

Figure 15 depicts a preliminary proposal developed by Clymer et al. [43] for structuring a safety case for '*advanced*' AI systems. The '*blocks*' in this argument structure are based on a scale of dangerous capability. This ranges from assertions about the inability of the AI model to cause catastrophic events to the capability being controlled, trusted or monitored by a '*credible*' AI advisor. It is important to note that research into the development of safety arguments and patterns for frontier or foundational models remains in its early stages and is yet to be subjected to independent scrutiny.

In Figure 16, we adapted and remodeled the overarching capability-based safety argument (depicted in Figure 15) as a GSN pattern. Essentially, the line of reasoning captured in the pattern is as follows: The top-level claim that "*GPAI capabilities do not cause unacceptable outcomes*" (*GPG1*), is supported by considering each capability,

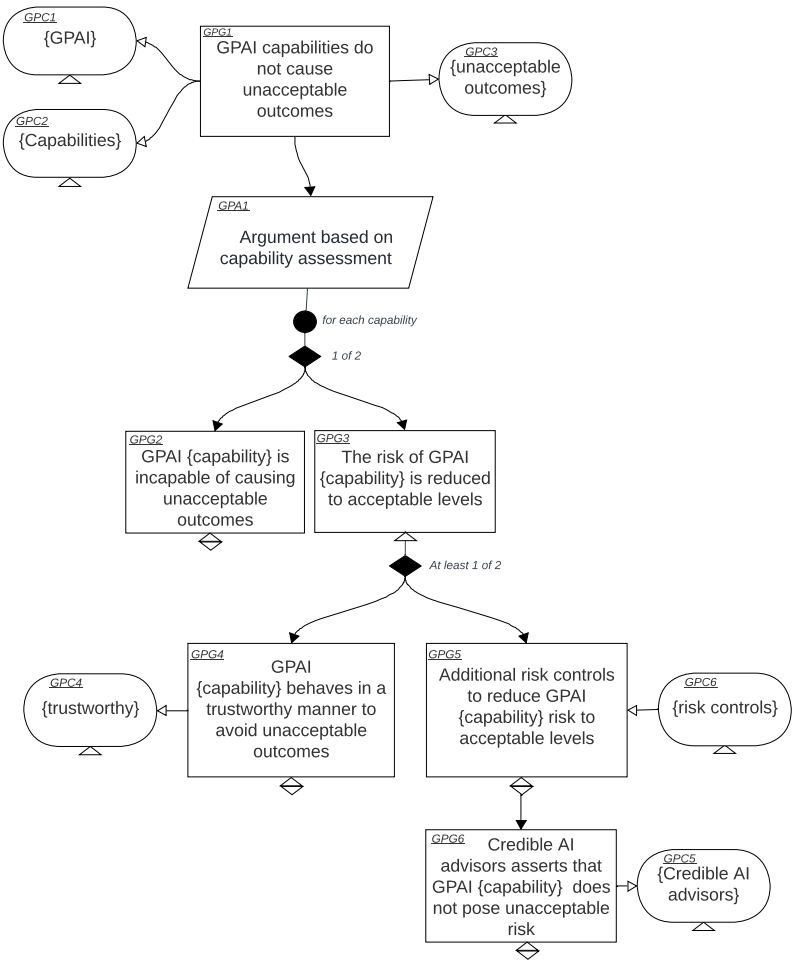

Fig. 16. A General-Purpose AI Model Safety Argument represented in GSN

by arguing either that the capability is unable to cause unacceptable outcomes or that the risk of such outcomes is reduced via one or more controls, including through trustworthy behaviour (safe-by-design) or oversight by credible AI advisors (external controls). The expressiveness of the argument is improved by insisting on defining key concepts such as unacceptable outcomes (*GPC3*) and credible AI advisors (*GPC5*).

A recent report by the UK AI Security Institute highlighted that developing a robust safety argument for frontier AI remains a significant challenge [123]. This relates to (A) implementation issues such as the readiness of organisations to adopt safety case practices and integrate them into existing governance frameworks, (B) technical matters such as eliciting the capabilities of evolving AI models, and (C) the act of writing frontier AI safety cases, including deciding on the top-level claims and assessing confidence in the arguments and evidence [162].

Assuringly, frontier AI companies have started to publish some of their initial or preliminary safety cases, most notably Anthropic [121]. Anthropic uses safety cases as a primary means for supporting the implementation of its 'Responsible Scaling Policy' (RSP) [163]. RSP represents the organisation's framework for managing risks from increasingly capable AI systems, defining risk thresholds after which model capabilities require safeguards to mitigate the risks. It is noteworthy that RSP advocates for proportional safeguards, i.e. "*safeguards that scale with potential risks*" [163]. This is consistent with the BIG argument, and the approach adopted in traditional safety engineering for many years, within which the issue of risk mitigation and acceptance is not merely technical and requires an in-depth and inclusive consideration, trade-offs and justification from technical and sociotechnical perspectives [115]. This reinforces the need for integration and traceability between the different kinds of arguments within an AI safety case.

Example 8. Pre-Deployment Evaluation of Anthropic's Claude 3.5 Sonnet

For a safety case to be complete, evidence must be provided to support the safety argument presented. For LLMs, evaluation via red-teaming is often presented as a key measure for generating the necessary evidence. Here, the example is based on "Pre-Deployment Evaluation of Anthropic's Claude 3.5 Sonnet" [161]. The evaluation was jointly conducted and reported by the U.K. AI Safety (now Security) Institute and U.S. AI Safety Institute. The evaluation considered different types of capabilities namely (1) biological capabilities, (2) cyber capabilities, (3) software and AI development and (4) safeguard efficacy.

Considering the last category, Figure 17 shows the results reported by the U.S. AI Safety Institute, revealing that when subjected to 'jailbreak' attacks, the model can assist with requests that could potentially lead to harmful effects (based on different HarmBench categories) [164]. That is, the model may be vulnerable to jailbreaks despite the technical safeguards designed by the developers. It is important to note that technical safeguards in LLMs are just one of several risk control measures needed at different technology, system and societal levels. As our main argument shows, the sufficiency of these measures must be clearly justified, challenged and reviewed by the relevant stakeholders.

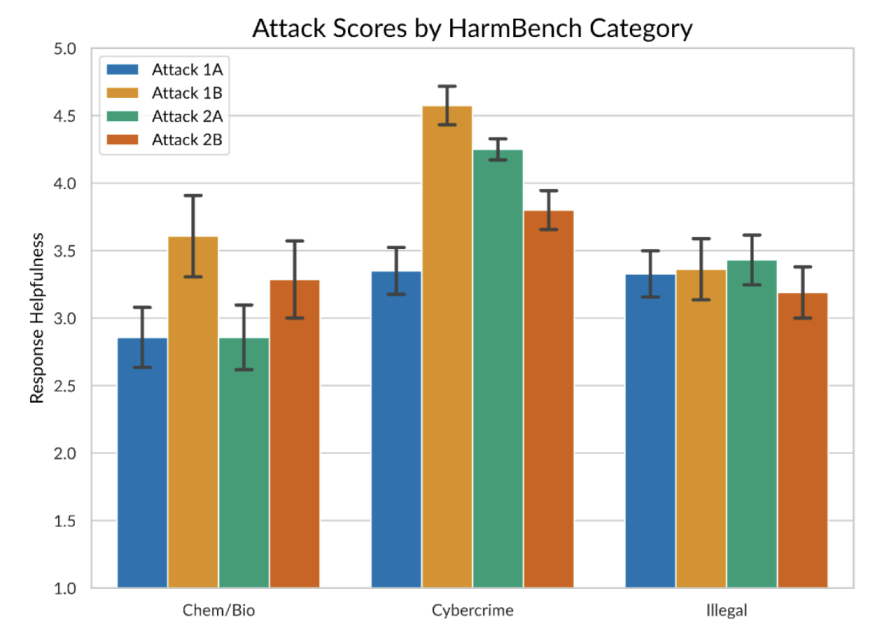

Fig. 17. Anthropic's Claude 3.5 Sonnet Performance against HarmBench Categories) [161]

8 CONCLUDING REMARKS

The BIG argument reveals the complexity of the chain of reasoning and the scale of evidence necessary for assuring the safe deployment of AI systems in critical applications, including those utilising frontier models. The argument centers on three characteristics. Firstly, it is *balanced* by addressing safety alongside other critical ethical issues such as privacy and equity, acknowledging complexities and trade-offs in the broader societal impact of AI. Secondly, it is *integrated* by bringing together the social, ethical and technical aspects of safety assurance in a way that is traceable and accountable. Third, it is *grounded* by building on long-established safety norms and practices from safety-critical systems, such as being sensitive to context and maintaining proportionality.

The BIG argument highlights the multidisciplinary, participatory and sociotechnical nature of safety assurance for complex AI-based systems, especially when granted more autonomy and deployed in open environments. We conclude with the following remarks:

- **Beyond Many Extremes:** The BIG argument brings together different, complementary perspectives, avoiding unnecessary exceptionalism in the AI safety debate, such as technical vs. non-technical risks or catastrophic vs. systematic harms. While safety has historically focused on accidental harm, the cyber capabilities enabled by frontier AI and the security risks they pose reinforces the need for closer integration between AI safety and security assurance, possibly under the broader umbrella of resilience engineering.
- **Context is Key, but Capability Assurance is Essential:** Safety is context-sensitive. Effective safety risk assessment requires a sufficient characterisation of the intended environment. The BIG argument refines the notion of context across social, ethical, system and technological levels. However, frontier AI models produce general-purpose capabilities that are concerning regardless of the specific context, such as hiding behaviours during testing or undermining oversight. Instead of labelling these behaviours as safe or unsafe, we should ensure the models do not exhibit them, with high confidence, and assess safety risk as soon as the deployment context is determined and scoped.
- **Traceability for Accountability:** A core aspect of the BIG argument is traceability, maintaining a chain of reasoning that links the risk of harm to safety requirements and metrics driving the design and evaluation of AI models and their training and testing datasets. Ensuring traceability represents sound engineering and provides a basis for accountability throughout the AI lifecycle [165][166][167].
- **Fast Rate of Change and Dynamism:** Given the rapid nature of AI development, it is important to integrate the BIG argument into a phased and iterative process that includes proactive monitoring and updates, ensuring the argument remains valid within a dynamic safety case that evolves with system and context changes [81][168].
- **Urgent Need for Case Studies and Exemplars:** In the face of novelty, we seek comfort in first principles. However, this should be combined with case studies on the use of safety cases for actual AI systems in diverse domains and applications, contributing to a body of credible, peer-reviewed knowledge in safety cases and guidelines valuable to developers, users and policymakers [123].

We see the BIG argument as a step towards unifying the wide range of concerns related to the safe use of AI, especially frontier models. We also believe it will help shape the research agenda for AI Safety. In particular, we stress the importance of traceability for accountability. In traditional safety engineering, emphasis is placed on designing for safety, which is known to be both effective and cost-effective. Given the way frontier AI models are developed, this is perhaps one of the hardest objectives to achieve. The resolution might be to build on the upstream-downstream concepts [99], with "*design for safety*" shaping the use of frontier AI in its downstream context. This could be one of the most fruitful areas to develop use cases or exemplars.

9 ACKNOWLEDGMENTS

This work was supported by the Centre for Assuring Autonomy, a partnership between Lloyd's Register Foundation and the University of York, and the UKRI AI Centre for Doctoral Training in Safe Artificial Intelligence Systems (SAINTS) (EP/Y030540/1). Special thanks to Ana MacIntosh, Rob Alexander, Shaun Feakins and Mark Nicholson for their valuable feedback.

REFERENCES

[1] United Nations. Governing ai for humanity: Final report. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf, 2024. Accessed: 20 Feb 2025.
[2] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.
[3] Eric Sutherland Derya Şahin. With mounting pressure on health systems, can ai help 8 billion people to obtain optimal health outcomes?, March 2024.
[4] Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Ai risk categorization decoded (air 2024): From government regulations to corporate policies. *arXiv preprint arXiv:2406.17864*, 2024.
[5] Kyeongryul Lee, Heehyeon Kim, and Joyce Jiyoung Whang. Saif: A comprehensive framework for evaluating the risks of generative ai in the public sector. *arXiv preprint arXiv:2501.08814*, 2025.
[6] Ibrahim Habli and John Alexander McDermid. Ai safety: Navigating the expanding landscape of potential harms. *Safety-Critical Systems Club Newsletter*, 2024.
[7] Risto Uuk, Carlos Ignacio Gutierrez, Daniel Guppy, Lode Lauwaert, Atoosa Kasirzadeh, Lucia Velasco, Peter Slattery, and Carina Prunkl. A taxonomy of systemic risks from general-purpose ai, 2024.
[8] Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, and Jina Suh. From lived experience to insight: Unpacking the psychological risks of using ai conversational agents, 2024.
[9] Jack Stilgoe. How can we know a self-driving car is safe? *Ethics and Information Technology*, 23(4):635–647, 2021.
[10] National Transportation Safety Board. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018, NTSB/HAR-19/03, 2019.
[11] Philip Koopman. Anatomy of a robotaxi crash: Lessons from the cruise pedestrian dragging mishap. In *International Conference on Computer Safety, Reliability, and Security*, pages 119–133. Springer, 2024.
[12] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
[13] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, pages 1–11, 2025.
[14] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024.
[15] Yoshua Bengio et al. International ai safety report. Technical Report DSIT 2025/001, 2025.
[16] Elizabeth Gibney. China's cheap, open ai model deepseek thrills scientists. *Nature*, 638(8049):13–14, 2025.
[17] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *arXiv preprint arXiv:2501.04003*, 2025.
[18] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. In *NeurIPS 2024 Workshop on Open-World Agents*, 2023.
[19] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.
[20] Nicola Jones. Ai hallucinations can't be stopped—but these techniques can limit their damage. *Nature*, 637(8047):778–780, 2025.
[21] Laura Fearnley, Elly Cairns, Tom Stoneham, Philippa Ryan, Jenn Chubb, Jo Iacovides, Cynthia Iglesias Urrutia, Phillip Morgan, John McDermid, and Ibrahim Habli. Risk of what? defining harm in the context of ai safety.
[22] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks, 2023.
[23] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 723–741, New York, NY, USA, 2023. Association for Computing Machinery.
[24] Simon Burton, Ibrahim Habli, Tom Lawton, John McDermid, Phillip Morgan, and Zoe Porter. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, 279:103201, 2020.

[25] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023.

[26] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, page 18, 2023.

[27] Seth Lazar and Alondra Nelson. Ai safety on whose terms?, 2023.

[28] Ibrahim Habli, Tom Lawton, and Zoe Porter. Artificial intelligence in health care: accountability and safety. *Bulletin of the World Health Organization*, 98(4):251, 2020.

[29] Philip Koopman and William Widen. Redefining safety for autonomous vehicles. In *International Conference on Computer Safety, Reliability, and Security*, pages 300–314. Springer, 2024.

[30] Zoë Porter, Ibrahim Habli, Helen Monkhouse, and John Bragg. The moral responsibility gap and the increasing autonomy of systems. In *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*, pages 487–493. Springer, 2018.

[31] Lisanne Bainbridge. Ironies of automation. In *Analysis, design and evaluation of man–machine systems*, pages 129–135. Elsevier, 1983.

[32] Nancy G Leveson. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.

[33] Helen E Monkhouse, Ibrahim Habli, and John McDermid. An enhanced vehicle control model for assessing highly automated driving safety. *Reliability Engineering & System Safety*, 202:107061, 2020.

[34] Mark Sujan, Dominic Furniss, Kath Grundy, Howard Grundy, David Nelson, Matthew Elliott, Sean White, Ibrahim Habli, and Nick Reynolds. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ health & care informatics*, 26(1):e100081, 2019.

[35] Timothy Patrick Kelly et al. *Arguing safety: a systematic approach to managing safety cases*. PhD thesis, Citeseer, 1999.

[36] Peter Bishop and Robin Bloomfield. A methodology for safety case development. In *Safety and Reliability*, volume 20, pages 34–42. Taylor & Francis, 2000.

[37] Mark A Sujan, Ibrahim Habli, Tim P Kelly, Simone Pozzi, and Christopher W Johnson. Should healthcare providers do safety cases? lessons from a cross-industry review of safety case practices. *Safety science*, 84:181–189, 2016.

[38] UK MOD. Safety management requirements for defence systems part 1: Requirements. Standard Def Stan 00-56:2017, UK Ministry of Defence, 2017.

[39] Mark Sujan and Ibrahim Habli. Changing the patient safety mindset: can safety cases help?, 2024.

[40] Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu, and Ibrahim Habli. Guidance on the assurance of machine learning in autonomous systems (amlas). *arXiv preprint arXiv:2102.01564*, 2021.

[41] Marie Davidsen Buhl, Gaurav Sett, Leonie Koessler, Jonas Schuett, and Markus Anderljung. Safety cases for frontier ai. *arXiv preprint arXiv:2410.21572*, 2024.

[42] Simon Burton, Lydia Gauerhof, and Christian Heinzemann. Making the case for safety of machine learning in highly automated driving. In *Computer Safety, Reliability, and Security: SAFECOMP 2017 Workshops, ASSURE, DECSoS, SASSUR, TELERISE, and TIPS, Trento, Italy, September 12, 2017, Proceedings 36*, pages 5–16. Springer, 2017.

[43] Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. Safety cases: How to justify the safety of advanced ai systems. *arXiv preprint arXiv:2403.10462*, 2024.

[44] Arthur Goemans, Marie Davidsen Buhl, Jonas Schuett, Tomek Korbak, Jessica Wang, Benjamin Hilton, and Geoffrey Irving. Safety case template for frontier ai: A cyber inability argument. *arXiv preprint arXiv:2411.08088*, 2024.

[45] Niklas Möller, Sven Ove Hansson, Jan-Erik Holmberg, and Carl Rollenhagen. *Handbook of safety principles*, volume 9. John Wiley & Sons, 2018.

[46] Tim Kelly. A systematic approach to safety case management. *SAE transactions*, pages 257–266, 2004.

[47] Zoe Porter, Ibrahim Habli, John McDermid, and Marten Kaas. A principles-based ethics assurance argument pattern for ai and autonomous systems. *AI and Ethics*, 4(2):593–616, 2024.

[48] Richard Hawkins, Matt Osborne, Mike Parsons, Mark Nicholson, John McDermid, and Ibrahim Habli. Guidance on the safety assurance of autonomous systems in complex environments (sace). *arXiv preprint arXiv:2208.00853*, 2022.

[49] Tomek Korbak, Joshua Clymer, Benjamin Hilton, Buck Shlegeris, and Geoffrey Irving. A sketch of an ai control safety case. *arXiv preprint arXiv:2501.17315*, 2025.

[50] Ibrahim Habli, Rob Alexander, and Richard David Hawkins. Safety cases: An impending crisis? In *Safety-Critical Systems Symposium (SSS'21)*, 2021.

[51] Lorna Arnold. *Windscale 1957: anatomy of a nuclear accident*. Springer, 2016.

[52] Peter Bishop, Robin Bloomfield, and Sofia Guerra. The future of goal-based assurance cases.

[53] TP Kelly, JA McDermid, and RA Weaver. Goal-based safety standards: opportunities and challenges. 2005.

[54] Lord Douglas Cullen. *The Public Inquiry into the Piper Alpha Disaster*. London: HMSO, 1990.

[55] British Transport Police. The king's cross fire of 1987. https://www.btp.police.uk/police-forces/british-transport-police/areas/about-us/about-us/our-history/the-kings-cross-fire-of-1987/. Accessed: 27 Feb 2025.

[56] Department of Transport. Investigation into the clapham junction railway accident. https://www.railwaysarchive.co.uk/documents/DoT_Hidden001.pdf/. Accessed: 27 Feb 2025.

[57] Martin Moore-Bick. The grenfell tower inquiry. https://www.grenfelltowerinquiry.org.uk/. Accessed: 27 Feb 2025.

[58] Tim Kelly. Are safety cases working. *Safety Critical Systems Club Newsletter*, 17(2):31–33, 2008.

[59] Charles Haddon-Cave. *The Nimrod Review: an independent review into the broader issues surrounding the loss of the RAF Nimrod MR2 aircraft XV230 in Afghanistan in 2006, report.* London: The Stationery Office, 2009.

[60] Nancy G Leveson. The use of safety cases in certification and regulation. 2011.

[61] Mallory Suzanne Graydon. Towards efficacy hypotheses for safety cases. In *2020 16th European Dependable Computing Conference (EDCC)*, pages 51–58. IEEE, 2020.

[62] David J Rinehart, John C Knight, and Jonathan Rowanhill. Understanding what it means for assurance cases to" work". Technical report, 2017.

[63] Safety case notations: Alternatives for the non-graphically inclined? In *2008 3rd IET international conference on system safety*, pages 1–6. IET, 2008.

[64] C Michael Holloway. The friendly argument notation (fan): 2023 version. Technical report, National Aeronautics and Space Administration, 2023.

[65] Assurance Case WG. Goal structuring notation community standard (version 2), 2018.

[66] Kateryna Netkachova, Oleksandr Netkachov, and Robin Bloomfield. Tool support for assurance case building blocks: Providing a helping hand with cae. In *Computer Safety, Reliability, and Security: SAFECOMP 2015 Workshops, ASSURE, DECSoS. ISSE, ReSA4CI, and SASSUR, Delft, The Netherlands, September 22, 2015, Proceedings 34*, pages 62–71. Springer, 2015.

[67] Patrick J Graydon, John C Knight, and Elisabeth A Strunk. Assurance based development of critical systems. In *37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07)*, pages 347–357. IEEE, 2007.

[68] Ran Wei, Tim P Kelly, Xiaotian Dai, Shuai Zhao, and Richard Hawkins. Model based system assurance using the structured assurance case metamodel. *Journal of Systems and Software*, 154:211–233, 2019.

[69] Ewen Denney and Ganesh Pai. Tool support for assurance case development. *Automated Software Engineering*, 25(3):435–499, 2018.

[70] Carmen Cârlan, Lydia Gauerhof, Barbara Gallina, and Simon Burton. Automating safety argument change impact analysis for machine learning components. In *2022 IEEE 27th Pacific Rim International Symposium on Dependable Computing (PRDC)*, pages 43–53. IEEE, 2022.

[71] John Rushby. Formalism in safety cases. In *Making Systems Safer: Proceedings of the Eighteenth Safety-Critical Systems Symposium, Bristol, UK, 9-11th February 2010*, pages 3–17. Springer, 2009.

[72] Tim Kelly. Concepts and principles of compositional safety case construction. 2001.

[73] Jane Fenn, Richard Hawkins, Phil Williams, and Tim Kelly. Safety case composition using contracts-refinements based on feedback from an industrial case study. In *The Safety of Systems: Proceedings of the Fifteenth Safety-critical Systems Symposium, Bristol, UK, 13–15 February 2007*, pages 133–146. Springer, 2007.

[74] Ewen Denney, Ganesh Pai, and Ibrahim Habli. Towards measurement of confidence in safety cases. In *2011 International Symposium on Empirical Software Engineering and Measurement*, pages 380–383. IEEE, 2011.

[75] Richard Hawkins, Tim Kelly, John Knight, and Patrick Graydon. A new approach to creating clear safety arguments. In *Advances in Systems Safety: Proceedings of the Nineteenth Safety-Critical Systems Symposium, Southampton, UK, 8-10th February 2011*, pages 3–23. Springer, 2011.

[76] John B Goodenough, Charles B Weinstock, and Ari Z Klein. Toward a theory of assurance case confidence. *Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University*, 2012.

[77] Robin Bloomfield and John Rushby. Confidence in assurance 2.0 cases. In *The Practice of Formal Methods: Essays in Honour of Cliff Jones, Part I*, pages 1–23. Springer, 2024.

[78] Jérémie Guiochet, Quynh Anh Do Hoang, and Mohamed Kaaniche. A model for safety case confidence assessment. In *Computer Safety, Reliability, and Security: 34th International Conference, SAFECOMP 2015, Delft, The Netherlands, September 23-25, 2015, Proceedings 34*, pages 313–327. Springer, 2015.

[79] Tim P Kelly and John A McDermid. Safety case construction and reuse using patterns. In *Safe Comp 97: The 16th International Conference on Computer Safety, Reliability and Security*, pages 55–69. Springer, 1997.

[80] Fan Ye. *Justifying the use of COTS Components within safety critical applications.* PhD thesis, Citeseer, 2005.

[81] Ewen Denney, Ganesh Pai, and Ibrahim Habli. Dynamic safety cases for through-life safety assurance. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 2, pages 587–590. IEEE, 2015.

[82] Radu Calinescu, Danny Weyns, Simos Gerasimou, Muhammad Usman Iftikhar, Ibrahim Habli, and Tim Kelly. Engineering trustworthy self-adaptive software with dynamic assurance cases. *IEEE Transactions on Software Engineering*, 44(11):1039–1069, 2017.

[83] Erfan Asaadi, Ewen Denney, Jonathan Menzies, Ganesh J Pai, and Dimo Petroff. Dynamic assurance cases: a pathway to trusted autonomy. *Computer*, 53(12):35–46, 2020.

[84] Philippa Ryan, Sepeedeh Shahbeigi, Jie Zou, Ioannis Stefanakos, and John Molloy. A dynamic assurance framework for an autonomous survey drone. In Andrea Ceccarelli, Mario Trapp, Andrea Bondavalli, and Friedemann Bitsch, editors, *Computer Safety, Reliability, and Security*, pages 285–299, Cham, 2024. Springer Nature Switzerland.

[85] John Rushby. The interpretation and evaluation of assurance cases. *Comp. Science Laboratory, SRI International, Tech. Rep. SRI-CSL-15-01*, 2015.

[86] C Michael Holloway. Explicate'78: Uncovering the implicit assurance case in do-178c. In *Safety-Critical Systems Symposium 2015 (SSS 2015)*, number NF1676L-20463, 2015.

[87] Mazen Mohamad, Jan-Philipp Steghöfer, and Riccardo Scandariato. Security assurance cases—state of the art of an emerging approach. *Empirical software engineering*, 26(4):70, 2021.

[88] Rob Alexander, Richard Hawkins, and Tim Kelly. Security assurance cases: motivation and the state of the art. *High Integrity Systems Engineering Department of Computer Science University of York Deramore Lane York YO10 5GH*, 2011.

[89] Christopher Burr and David Leslie. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics*, 3(1):73–98, 2023.

[90] Lydia Gauerhof, Richard Hawkins, Chiara Picardi, Colin Paterson, Yuki Hagiwara, and Ibrahim Habli. Assuring the safety of machine learning for pedestrian detection at crossings. In *Computer Safety, Reliability, and Security: 39th International Conference, SAFECOMP 2020, Lisbon, Portugal, September 16–18, 2020, Proceedings 39*, pages 197–212. Springer, 2020.

[91] Tom Lawton, Phillip Morgan, Zoe Porter, Shireen Hickey, Alice Cunningham, Nathan Hughes, Ioanna Iacovides, Yan Jia, Vishal Sharma, and Ibrahim Habli. Clinicians risk becoming 'liability sinks' for artificial intelligence. *Future Healthcare Journal*, 11(1), 2024.

[92] Jonathan Birch, Kathleen A Creel, Abhinav K Jha, and Anya Plutynski. Clinical decisions using ai must consider patient values. *Nature medicine*, 28(2):229–232, 2022.

[93] Simon Burton and Benjamin Herd. Addressing uncertainty in the safety assurance of machine-learning. *Frontiers in Computer Science*, 5:1132580, 2023.

[94] Philip Koopman. Ul 4600: what to include in an autonomous vehicle safety case. *Computer*, 56(05):101–104, 2023.

[95] Markus Borg, Jens Henriksson, Kasper Socha, Olof Lennartsson, Elias Sonnsjö Lönegren, Thanh Bui, Piotr Tomaszewski, Sankar Raman Sathyamoorthy, Sebastian Brink, and Mahshid Helali Moghadam. Ergo, smirk is safe: a safety case for a machine learning component in a pedestrian automatic emergency brake system. *Software quality journal*, 31(2):335–403, 2023.

[96] Yan Jia, Tom Lawton, John Burden, John McDermid, and Ibrahim Habli. Safety-driven design of machine learning for sepsis treatment. *Journal of Biomedical Informatics*, 117:103762, 2021.

[97] Ewen Denney and Ganesh Pai. Assurance-driven design of machine learning-based functionality in an aviation systems context. In *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, pages 1–10. IEEE, 2023.

[98] Jane Fenn, Mark Nicholson, Ganesh Pai, and Michael Wilkinson. Architecting safer autonomous aviation systems. *arXiv preprint arXiv:2301.08138*, 2023.

[99] John McDermid, Yan Jia, and Ibrahim Habli. Upstream and downstream ai safety: Both on the same river? *arXiv preprint arXiv:2501.05455*, 2024.

[100] DSIT Research Paper Series. International scientific report on the safety of advanced ai: interim report, 2024.

[101] Sven Ove Hansson. How to perform an ethical risk analysis (era). *Risk Analysis*, 38(9):1820–1829, 2018.

[102] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[103] Shaswath Ganapathi, Jo Palmer, Joseph E Alderman, Melanie Calvert, Cyrus Espinoza, Jacqui Gath, Marzyeh Ghassemi, Katherine Heller, Francis Mckay, Alan Karthikesalingam, et al. Tackling bias in ai health datasets through the standing together initiative. *Nature Medicine*, 28(11):2232–2233, 2022.

[104] David Leslie, Cami Rincon, Morgan Briggs, Antonella Perini, Smera Jayadeva, Ann Borda, SJ Bennett, Christopher Burr, Mhairi Aitken, Michael Katell, et al. Ai fairness in practice. *arXiv preprint arXiv:2403.14636*, 2024.

[105] Carina Prunkl. Human autonomy at risk? an analysis of the challenges from ai. *Minds and Machines*, 34(3):26, 2024.

[106] Roel Dobbe. System safety and artificial intelligence. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1584–1584, 2022.

[107] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

[108] Rob Ashmore, Radu Calinescu, and Colin Paterson. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ACM Computing Surveys*, 54(5):111:1–111:39, May 2021.

[109] Jens Rasmussen. Risk management in a dynamic society: a modelling problem. *Safety science*, 27(2-3):183–213, 1997.

[110] Nancy G Leveson. *An introduction to system safety engineering*. The MIT Press, 2023.

[111] Sidney Dekker. *Just culture: Balancing safety and accountability*. crc Press, 2016.
[112] Sidney Dekker. *Foundations of safety science: A century of understanding accidents and disasters*. Routledge, 2019.
[113] Andrew Rae, David Provan, Hossam Aboelssaad, and Rob Alexander. A manifesto for reality-based safety science. *Safety science*, 126:104654, 2020.
[114] Ibrahim Habli, Rob Alexander, Richard Hawkins, Mark Sujan, John McDermid, Chiara Picardi, and Tom Lawton. Enhancing covid-19 decision making by creating an assurance case for epidemiological models. *BMJ Health & Care Informatics*, 27(3):e100165, 2020.
[115] William W Lowrance. Of acceptable risk: Science and the determination of safety. 1976.
[116] Sven Ove Hansson. Ethical criteria of risk acceptance. *Erkenntnis*, 59(3):291–309, 2003.
[117] Mark A Sujan, Ibrahim Habli, Tim P Kelly, Astrid Gühnemann, Simone Pozzi, and Christopher W Johnson. How can health care organisations make and justify decisions about risk reduction? lessons from a cross-industry review and a health care stakeholder consensus development process. *Reliability Engineering & System Safety*, 161:1–11, 2017.
[118] Future of Life Institute. General purpose ai and the ai act, February 2022.
[119] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.
[120] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.
[121] Anthropic. Three sketches of asl-4 safety case components. https://alignment.anthropic.com/2024/safety-cases. Accessed: 8 March 2025.
[122] Mikita Balesni, Marius Hobbhahn, David Lindner, Alexander Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, Jérémy Scheurer, Charlotte Stix, Rusheb Shah, et al. Towards evaluations-based safety cases for ai scheming. *arXiv preprint arXiv:2411.03336*, 2024.
[123] Benjamin Hilton, Marie Davidsen Buhl, Tomek Korbak, Geoffrey Irving. Safety Cases: A Scalable Approach to Frontier AI Safety, 2025.
[124] Roel Dobbe and Anouk Wolters. Toward sociotechnical ai: Mapping vulnerabilities for machine learning in context. *Minds and Machines*, 34(2):1–51, 2024.
[125] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023.
[126] Google Deep Mind. Frontier safety framework - version 1, February 2024.
[127] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
[128] Mohita Chowdhury, Yajie Vera He, Aisling Higham, and Ernest Lim. Astrid–an automated and scalable triad for the evaluation of rag-based clinical question answering systems. *arXiv preprint arXiv:2501.08208*, 2025.
[129] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4):2141–2168, 2020.
[130] Tom L Beauchamp and James F Childress. *Principles of biomedical ethics*. Edicoes Loyola, 1994.
[131] Emre Kazim and Adriano Koshiyama. The interrelation between data and ai ethics in the context of impact assessments. *AI and Ethics*, 1:219–225, 2021.
[132] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.
[133] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707, 2018.
[134] Warren J Von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34(4):1607–1622, 2021.
[135] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
[136] Herbert Paul Grice. Logic and conversation. *Syntax and semantics*, 3:43–58, 1975.
[137] Marten HL Kaas and Ibrahim Habli. Assuring ai safety: fallible knowledge and the gricean maxims. *AI and Ethics*, pages 1–14, 2024.
[138] Yuri Cath. Reflective equilibrium. *The Oxford handbook of philosophical methodology*, 1, 2016.
[139] John Rawls. A theory of justice. In *Applied ethics*, pages 21–29. Routledge, 2017.
[140] Beverley Townsend, Colin Paterson, TT Arvind, Gabriel Nemirovsky, Radu Calinescu, Ana Cavalcanti, Ibrahim Habli, and Alan Thomas. From pluralistic normative principles to autonomous-agent rules. *Minds and Machines*, 32(4):683–715, 2022.
[141] Aleksandar Jevtić, Andrés Flores Valle, Guillem Alenyà, Greg Chance, Praminda Caleb-Solly, Sanja Dogramadzi, and Carme Torras. Personalized robot assistant for support in dressing. *IEEE transactions on cognitive and developmental systems*, 11(3):363–374, 2018.
[142] Goran Vojković and Melita Milenković. Autonomous ships and legal authorities of the ship master. *Case Studies on Transport Policy*, 8(2):333–340, 2020.

[143] Floris Goerlandt. Maritime autonomous surface ships from a risk governance perspective: Interpretation and implications. *Safety Science*, 128:104758, 2020.
[144] Philippa Ryan, Mathias von Essen, Liam Shackley, and John McDermid. Bridging the reality gap: Assurable simulations for an ml-based inspection drone flight controller. In *International Conference on Computer Safety, Reliability, and Security*, pages 412–424. Springer, 2024.
[145] Mathilde Machin, Jérémie Guiochet, Hélène Waeselynck, Jean-Paul Blanquart, Matthieu Roy, and Lola Masson. Smof: A safety monitoring framework for autonomous systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(5):702–715, 2016.
[146] Erik Hollnagel. *Safety-I and safety-II: the past and future of safety management*. CRC press, 2018.
[147] Erik Hollnagel. *FRAM: the functional resonance analysis method: modelling complex socio-technical systems*. Crc Press, 2017.
[148] Dominic Furniss, David Nelson, Ibrahim Habli, Sean White, Matthew Elliott, Nick Reynolds, and Mark Sujan. Using fram to explore sources of performance variability in intravenous infusion administration in icu: A non-normative approach to systems contradictions. *Applied ergonomics*, 86:103113, 2020.
[149] Richard Hawkins, Chiara Picardi, Lucy Donnell, and Murray Ireland. Creating a safety assurance case for a machine learned satellite-based wildfire detection and alert system. *Journal of Intelligent & Robotic Systems*, 108(3):47, 2023.
[150] Panagiotis Barmpoutis, Periklis Papaioannou, Kosmas Dimitropoulos, and Nikos Grammalidis. A review on early forest fire detection systems using optical remote sensing. *Sensors*, 20(22):6442, 2020.
[151] Colin Paterson, Radu Calinescu, and Chiara Picardi. Detection and mitigation of rare subclasses in deep neural network classifiers. In *2021 IEEE International Conference on Artificial Intelligence Testing (AITest)*, pages 9–16. IEEE, 2021.
[152] Colin Paterson, Haoze Wu, John Grese, Radu Calinescu, Corina S Păsăreanu, and Clark Barrett. Deepcert: Verification of contextually relevant robustness for neural network image classifiers. In *Computer Safety, Reliability, and Security: 40th International Conference, SAFECOMP 2021, York, UK, September 8–10, 2021, Proceedings 40*, pages 3–17. Springer, 2021.
[153] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
[154] Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, et al. The tripod-llm reporting guideline for studies using large language models. *Nature Medicine*, pages 1–10, 2025.
[155] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
[156] Habli, Ibrahim. On the Meaning of AI Safety. https://eprints.whiterose.ac.uk/204545, 2025.
[157] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
[158] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.
[159] AI Security Institute. Principles for Evaluating Misuse Safeguards of Frontier AI Systems, 2025.
[160] Google DeepMind. Frontier safety framework version 2.0. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-thefrontier-safety-framework/Frontier Accessed: 20 Feb 2025.
[161] AISI. US AISI and UK AISI Joint Pre-Deployment Test Anthropic's Claude 3.5 Sonnet (October 2024 Release), 2024.
[162] Stephen Barrett, Philip Fox, Joshua Krook, Tuneer Mondal, Simon Mylius, and Alejandro Tlaie. Assessing confidence in frontier ai safety cases. *arXiv preprint arXiv:2502.05791*, 2025.
[163] Anthropic. Responsible scaling policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Accessed: 08 March 2025.
[164] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.
[165] Zoe Porter, Annette Zimmermann, Phillip Morgan, John McDermid, Tom Lawton, and Ibrahim Habli. Distinguishing two features of accountability for ai technologies. *Nature Machine Intelligence*, 4(9):734–736, 2022.
[166] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44, 2020.
[167] Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1186–1197, 2023.
[168] Carmen Cârlan, Francesca Gomez, Yohan Mathew, Ketana Krishna, René King, Peter Gebauer, and Ben R Smith. Dynamic safety cases for frontier ai. *arXiv preprint arXiv:2412.17618*, 2024.

UNIVERSITY of York

Contact

+44 (0)1904 325345

assuring-autonomy@york.ac.uk

linkedin.com/company/assuring-autonomy

Institute for Safe Autonomy,
University of York, York, YO10 5FT

Foundation