AI and Ethics
https://doi.org/10.1007/s43681-024-00490-x

ORIGINAL RESEARCH

Assuring AI safety: fallible knowledge and the Gricean maxims

Marten H. L. Kaas[1] · **Ibrahim Habli**[1]

Received: 2 January 2024 / Accepted: 2 May 2024

Abstract
In this paper we argue that safety claims, when justified by a safety case, are descriptive fallible knowledge claims. Even if the aim of a safety case was to justify infallible knowledge about the safety of a system, such infallible safety knowledge is impossible to attain in the case of AI-enabled systems. By their nature AI-enabled systems preclude the possibility of obtaining infallible knowledge concerning their safety or lack thereof. We suggest that one can communicate knowledge of an AI-enabled system's safety by structuring their exchange according to Paul Grice's Cooperative Principle which can be achieved via adherence to the Gricean maxims of communication. Furthermore, these same maxims can be used to evaluate the calibre of the exchange, with the aim being to ensure that communicating knowledge about an AI-enabled system's safety is of the highest calibre, in short, that the communication is relevant, of sufficient quantity and quality, and communicated perspicuously. The high calibre communication of safety claims to an epistemically diverse group of stakeholders is vitally important given the increasingly participatory nature of AI-enabled system design, development and assessment.

Keywords Safety assurance · Artificial intelligence · Epistemology · Gricean maxims · Safety communication

1 Introduction

A clinician, Clara, makes her way to see her last patient, Peter, before her shift ends. It has been a long day made even more difficult by the fact that most of her time is spent with patients whose cases are complicated and out of the ordinary. Routine but cognitively light tasks have been increasingly automated away through the use of autonomous AI-enabled technologies. Clara no longer conducts routine follow-up calls with patients, which were often pleasant and a welcome reprieve for her, and instead spends most of her day handling cases that an AI-enabled recommender system has flagged as "urgent." Peter's is one such case. Peter is a type-2 diabetic and was admitted to the hospital after complaining about chest pain. While Clara would normally meet with Peter alone, her hospital has recently acquired a clinical diagnostic support system named HIPPOCRATES that is supposed to enhance human diagnostic ability. After examining Peter, Clara is worried that his chest pain is symptomatic of an impending myocardial infarction, i.e., heart attack. HIPPOCRATES, in contrast, does not predict that Peter will experience a myocardial infarction within the next three months. While treatment decisions ultimately rest with Clara (and Peter of course), she is frustrated by the system's disagreement with her assessment. She is mindful of many different considerations; Peter's health, scarce hospital resources, her relatively new relationship with HIPPOCRATES, and the consequences of acting on her own judgement versus deferring to the recommendation of HIPPOCRATES. In her exhausted state, Clara recommends a standard treatment for Peter and sends him on his way.

This hypothetical scenario involving a clinician and a CDSS (clinical diagnostic support system) might seem farfetched, but AI-enabled systems already perform many of the tasks mentioned in the vignette above. AI-enabled systems can conduct routine follow up calls [29], can engage in triaging patients [9, 45], and can diagnose and treat illnesses like sepsis [25]. While there are numerous different issues that we might draw out and analyse from the above story, e.g., questions about the transparency of such an AI-enabled system, our focus is going to be on the issue of AI safety. When is it safe to deploy an AI-enabled system? Moreover, what does the claim that a given AI-enabled system is safe amount to? Is it a knowledge claim? Or just a claim about what one believes to be the case?

✉ Marten H. L. Kaas
marten.kaas@york.ac.uk

[1] Institute for Safe Autonomy, University of York, Deramore Lane, York YO10 5GH, UK

Published online: 15 May 2024

In this paper we will argue that safety claims about AI-enabled systems are claims of knowledge, i.e., about what one knows and not merely believes, if they are supported by an appropriately structured argument which is, in turn, justified by defeasible evidence. Indeed this is exactly the purpose of a safety case, to provide a clear, comprehensible, and defensible argument, supported by a body of evidence, that a system is acceptably safe to operate in a given context [28, 42]. The purpose of the safety case methodology, i.e., the process of producing a safety case, and the resultant safety case[1] is to supply justification for the knowledge claim that a given system is acceptably safe. This however is not *infallible* knowledge but rather *fallible* knowledge. This is because safety cases have their roots in the tradition of *informal argumentation*. The conclusions of or claims supported by informal arguments are rarely, if ever, established with certainty in the way that the conclusions of formal arguments are established.

This is particularly evident when considering the safety of a system vis-à-vis a safety case and the claim that the system is acceptably safe to operate. But even if the aim of a safety case, or informal argument in general, was to produce infallible knowledge, the nature of AI-enabled systems would preclude this possibility. In other words, *infallible knowledge claims* about AI-enabled systems, e.g., their behaviour, safety, etc., are unattainable in practice. These topics are taken up in sections two and three respectively. In section four, we attempt to answer the following question. Given claims about an AI-enabled system's safety (justified by a safety case), how best to communicate this knowledge, especially to an epistemically diverse group of stakeholders? Our novel contribution is to suggest that the Gricean maxims of cooperative communication can be used to evaluate the calibre of the communication between affected stakeholders.

2 Safety–critical systems

There are many different technological systems whose failure could result in loss of life, loss or significant damage to equipment and/or property, or damage to the environment. These safety–critical systems are particularly common in certain domains like healthcare, defence, aviation, and the petrochemical sector to name a few, but they are by no means limited to these domains [30, 41].

Safety is commonly conceptualised as freedom from harm [19]. But given that absolute freedom is rarely, if ever, possible for complex systems, definitions of safety tend to focus on the notion of risk, i.e. the likelihood and severity of harm [39]. This triggers necessary questions of acceptability of risk, by whom and given what else. It is important to note that intent matters. Harm here is unintended. It is typically due to error or complexity. If harm is deliberate, and it involves malice, then conceptually, the risk of harm falls within the realm of security and not safety, though both safety and security need to be considered in an integrated manner [3].

Here, safety is conceptualised as a *state*, i.e. a condition of the system in which it is free from harm. Other approaches, though not mutually exclusive, are more action-oriented, describing safety as the prevention or control of unacceptable or intolerable risk of harm. Recent safety science literature, under the umbrella of Safety II [22] or Resilience Engineering [23], emphasises a different perspective: safety is achieved through the adaptive capacity of the sociotechnical system to adjust its behaviour under both expected and unexpected conditions. It focuses on how "*things go right, rather than by preventing them from going wrong*" [22].

Regardless of the specific definition of, or perspective on, safety, AI-enabled functions are increasingly seen as standing in need of safety assurance either because their adoption raises questions about safety or because they are being integrated into safety–critical systems. But how is a system deemed "safe enough" to deploy? More importantly for our purposes, what kind of claim is one making when they state that a system is "safe enough" or, synonymously, "acceptably safe"?[2]

2.1 Safety assurance via safety cases

Before discussing what kind of claim a safety claim is, it is important to contextualise the practice of producing safety cases in order to assure the safety of a system. Safety practices have evolved significantly over the last fifty years [8]. Considerations of safety were initially, and unfortunately, reactive. The petrochemical, nuclear and railway domains for example are replete with accidents, many of which were catastrophic, that precipitated changes to safety practices.[3] It was largely only after accident investigations that changes to systems were made, if they were made at all, to ensure that similar accidents would not occur again in the future. Regulation in these domains was similarly reactive in the sense that manufacturers and operators had to meet specific standards and technical requirements specified by regulators who were not able to keep pace with technological innovations. The result was, for two main reasons, safety management

[1] This can sometimes be referred to as a safety report.

[2] For simplicity, our usage of the term 'safe' is also synonymous with the terms 'safe enough' and 'acceptably safe' unless otherwise specified.

[3] See (Sujan et al., 2012) for a brief chronological summary of significant events and their impact on safety regulation.

that was not fit for purpose [41]. First, this approach led to a culture of "box ticking" where the focus was more on compliance with standards and less on actually understanding and managing risks. Second, because the emphasis was on compliance with standards and regulations, this approach stifled innovation and hindered progress in industries driven by technological change.

In response to major accidents and changing economic environments (e.g., the privatisation of public industries) approaches to demonstrating the safety of a given system began to change. In addition to demonstrating compliance with applicable standards and requirements, current approaches to safety "require manufacturers and operators to demonstrate that they have adopted a thorough and systematic process to proactively understand the risks associated with their systems and control these risks appropriately" [41]. These duties can be fulfilled through the use of safety cases, i.e., appropriately structured arguments justified by defeasible evidence. Importantly, this is not to say that safety cases alone have led to improving safety practices. In addition to the adoption of safety cases there has been, for example, more proactive safety management in general as well as a more widespread safety culture that have also contributed to improving safety practices.

2.2 Safety claims

In its simplest form, a safety case is a clear, comprehensive and defensible argument that a system is acceptably safe to operate in a given context [28]. In what follows, our focus will be on the safety of AI-enabled systems unless otherwise specified. Additionally, we will primarily be referring to safety cases and safety case production as a monolithic enterprise. This however obscures some of the variation between the different schools of thought when it comes to safety case production [16]. There are even some inconsistencies that can arise if one equivocates on the meaning of "safety case," e.g., including voluminous technical details that interrupts or obscures the story of a system's safety because one believes that the purpose of a safety case is to show how safety requirements are satisfied through different levels of design [16]. While we do not commit ourselves to any one safety case school (which are not mutually exclusive, we do want to draw particular attention to two lines of thought that are pertinent for our upcoming discussion of AI safety, (1) that a safety case is used to document and *communicate* the story of a system's safety to diverse stakeholders (more on this in Sect. 4) including what it means for the system to be safe and how it achieves safety, and (2) that a safety case is used to establish confidence in safety claims, i.e., it is used to assure claims about safety [16]. There are also two major distinctions to note between a safety case and the safety case methodology. The former is an instantiated and compelling argument intended to support the claim that a given system is acceptably safe. The latter, appropriately, refers to the process by which one constructs or produces the safety case. There are different ways to present a safety case, e.g., images, text, bespoke notation, etc., and different methodologies one might use to produce the safety case. Caveats and clarifications aside, we turn now to consider what kind of claim is advanced in a safety case.

2.2.1 Safety as a descriptive claim

Claims advanced in a safety case about an AI-enabled system's safety, or any system's safety, are descriptive claims. They are about states of affairs. Let us consider the example of an AI-enabled extubation system that we will refer back to throughout this paper. In intensive care units (ICUs) patients may require invasive mechanical ventilation if they cannot breathe unaided. Intubation is the term used for the insertion of a tube into the trachea for such patients and extubation is the term used for the removal of the tube. An AI-enabled system can be used to predict patient readiness for extubation, a safety–critical task given the harmful consequences associated with both early and late weaning from mechanical ventilation [25]. To claim that this AI-enabled extubation system is acceptably safe to operate is not to say anything about what *ought* to be the case, i.e., something normative (e.g., we ought to deploy the system), but rather to say something about what *is* or will be the case, i.e., something descriptive (e.g., the system falsely predicted X percent of patients as ready for extubation in the test dataset). Note that while claims about an AI-enabled system's safety are descriptive, one might make certain normative claims about those descriptive claims. For example, we might claim that you should not believe the claim that the AI-enabled extubation system is acceptably safe.

It might however be objected that claims about an AI-enabled system's safety are inherently or implicitly normative. For example, it is relatively common to infer that one can or ought to do something because it is safe to do so. Utterances like, "Elevators are safe to ride in" or "It's safer to fly on a plane than drive in a car", seem to suggest that one ought to take the elevator or that one ought not be afraid of flying. Granting that interpretation of the above utterances, it is nevertheless possible to separate descriptive claims about safety from normative ones. That is, it is possible to separate the factual/descriptive dimension from the evaluative/normative dimension of safety. The former invariably revolves around physical, technical or measurable facts. For example one might claim that the elevator is safe enough because the steel cables supporting it have a certain tensile strength two orders of magnitude greater than the elevator and any load it might carry. It is in this descriptive dimension that factual or technical judgments dominate given their role in justifying

the claim that the system in question (e.g., the elevator) *is* safe enough. Normative claims in contrast revolve around the adequacy or evaluation of these physical facts. For example one might claim that an elevator ought not be considered safe enough unless the steel cables supporting it have a minimum tensile strength three orders of magnitude greater than the elevator and any load it might carry. Similarly, one might claim that different or additional tests ought to be used to re/evaluate the tensile strength of the steel cables.

In short, normative claims pertain to the safety requirements, e.g., whether they are sufficient or whether the proposed threshold that constitutes "safe enough" is appropriate, whereas descriptive claims pertain to the fulfilment of those requirements, e.g., whether it is the case that the requirements have been satisfied or the threshold of "safe enough" met. Importantly, when one makes a safety claim, e.g., that the elevator is safe enough or the AI-enabled extubation system is safe enough, the normative work has already been done; the value judgments were made when the safety requirements and threshold of "safe enough" were set.[4] It is, of course, always possible to revisit or question these requirements or threshold, which is the normative dimension of safety, but our point is that that kind of activity can be separated, at least conceptually, from the descriptive dimension of safety.

2.2.2 Safety as a knowledge claim

Claims advanced in a safety case about an AI-enabled system's safety, or again any system's safety, are knowledge claims. That is, one is justified in believing the truth (assuming the truth) of the claim that the system is acceptably safe to operate. But this can be developed further. For example, one might rightfully object that this rudimentary JTB (justified true belief) account of knowledge has serious flaws, one of which is its susceptibility to Gettier cases [12]. Gettier argued that it just might be by chance, for example, that one holds the justified true belief that the system described in the safety case is safe. Possessing a justified true belief therefore appears necessary but ultimately insufficient to secure a knowledge claim. Some additional requirement is required to ensure that the justification, the belief, or both is of the right sort to prevent cases of justified false beliefs from being counted as knowledge when epistemic luck is involved. But we are straying dangerously close to an analysis of epistemological issues stemming from Gettier cases that is well outside the scope of this paper. The only requirement that we will add here is that the justification be produced by a reliable process. This is the process reliabilist view according to which justification is a function of the reliability of the process or processes that cause one to form beliefs[5] that are true rather than false [14]. So one's belief that a given AI-enabled system is acceptably safe is justified and counts as knowledge just in case the process that caused one to form that belief, e.g., reading and understanding the safety test results, reliably produces true beliefs.

Importantly, though we can mitigate concerns about epistemic luck via the reliabilist justification requirement, there is still the possibility that our justified belief turns out to be false. The knowledge claim advanced in a safety case is therefore *fallible* knowledge. It is not *certain* knowledge, where certainty here is understood as *infallibility*, i.e., it could not have been false. This is distinct from other kinds of subjective certainty, e.g., incorrigibility about what I believe or feel at a particular moment, and distinct from other kinds of epistemic certainties, e.g., indubitability or indefeasibility about what I know or am justified in believing [37]. Put simply, fallibilism is the view that conjoins two strongly held intuitions that, on the one hand, we can make mistakes and sometimes be mistaken about things but that, on the other hand, we also have quite a bit of knowledge and can know things in spite of the mistakes we might make [36]. So claims advanced in a safety case about an AI-enabled system's safety are fallible knowledge claims because they can turn out, in unfortunate cases where a mistake has been made, to be false.[6]

2.3 Establishing safety claims

Claims about the safety of AI-enabled systems are descriptive fallible knowledge claims. When justified by a safety case, claims about the safety of an AI-enabled extubation system, for example, are more than mere belief. This

[4] Indeed value-laden judgments are inescapable and affect both the context of discovery and the context of evaluation. Roughly speaking, values can influence what domain or subject a person studies, can influence hypothesis formation, and the choice of evidence to be gathered (which is most relevant for the current discussion), all of which are generally part of the context of discovery. But values can also influence how a person interprets the evidence that has been gathered (which is, again, most relevant for the current discussion), the method of analysis employed, and the evidence's relation to the hypothesis and larger theoretical constructs, all of which are generally part of the context of evaluation [24]. So while in practice it may be difficult to separate the normative and descriptive dimensions of safety, e.g., because two individuals might interpret the same evidence as in/sufficient for meeting a given safety requirement or threshold, we can nonetheless conceptually separate the normative and descriptive dimensions.

[5] This can be generalised to include information in general. That is, some have argued that AI-enabled machines can produce "knowledge" because their outputs reliably lead to the formation of true "beliefs.".

[6] This can include mistakes arising from fallacious reasoning, some of which have already been documented in the context of safety cases [17].

knowledge is fallible however because the justification provided vis-à-vis the safety case relies on a certain type of argument, namely informal argumentation as opposed to formal argumentation. We turn now to briefly describe the differences between formal and informal arguments as well as how and why informal arguments leave room for the fallibility of their conclusions.

2.3.1 Formal arguments

As the name suggests, the idea of formal arguments arises from an intuition that arguments, and reason more broadly, ought to be systematically formalized. Reasoning and arguments, in short, ought to be thought of as a science whose object of study is logical relations and those laws and phenomena governed by logical relations [43]. The laws[7] include, for example, those concerning entailment and deduction, and the phenomena include, for example, the properties of validity and well-formedness.

2.3.1.1 Advantages and disadvantages One immense advantage of formal arguments is their ability to demonstrate the infallibility of a conclusion. Provided that the argument begins with premises or axioms that are true (or accepted as such), and provided that no mistakes were made in the ensuing inferences, the conclusion must be true. One stark example of this is Kurt Gödel's famous incompleteness theorem which demonstrated, beginning with accepted axioms, that there are certain true statements or conclusions that are impossible to state. Gödel's conclusion, that given a sufficiently complex logic there are certain conclusions that cannot be written in the language of that logic, hence their incompleteness, is definitive and *infallible*, i.e., it could not have been false.

A related advantage of formal arguments is their specificity. Natural language is fraught with ambiguity that can render even relatively simple sentences, let alone arguments, difficult to understand. Consider the following sentence. "The battery is empty, and the robot is not moving, or the patient is hungry." This sentence can be parsed in two different ways as the battery is empty and either the robot is not moving or the patient is hungry, or as either the battery is empty and the robot is not moving or the patient is hungry. In a formal argument, such a sentence might be represented by the string (A&(BVC)) or ((A&B)VC) respectively, resolving any ambiguities in interpretation.

In spite of their advantages, formal arguments also suffer from significant disadvantages. Perhaps the most serious of which is that formal arguments are largely disconnected from the way in which people normally argue and reason. The average person is often less concerned with the validity of a mathematical proof and more often concerned with assessing the reasons and arguments a colleague provides for using an AI-enabled extubation system, for example. Moreover, the average person is often less interested in reasoning about what is certainly the case (or not) and more often interested in reasoning about what is probably or likely to be the case (or not), as one might be when conversing with their physician about back pain they are experiencing. Conversations like these can certainly be instantiated using formal arguments, but the usefulness of doing so is dubious at best.

Formal arguments are similarly disconnected from the real world. By that we mean that formal arguments are largely disconnected from the messy, uncertain and dynamic realities that overwhelmingly dominate our human existence. Rarely, for instance, are there widely agreed upon axioms or starting points from which one can uncontroversially begin their natural argument. Such is the case with safety assurance generally, and for AI-enabled systems more specifically, where one can always, as we saw above, raise legitimate normative concerns/questions. Likewise, natural arguments are rarely presented using the (mostly) unambiguous language of symbolic logic but instead using natural language with all of its accompanying ambiguities, vagueness and complexities. Natural arguments between two (or more) persons are more of a dialectical reasoning process than a sterile evaluation of logical relations between different variables.

This brings us to a third disadvantage of formal arguments, and one that follows from the first two described, namely that formal arguments are ideal abstractions. From a safety perspective, this is a critical defect. While the idealized and abstract nature of formal arguments confers certain advantages, this is at the cost of neglecting potential crucial context-dependent information. Safety is not merely a technical property but increasingly a socio-technical property that depends not just on the system itself but on how people interact with the system. An AI-enabled extubation system could be acceptably safe to operate in an ideal world, but be unsafe to operate when one considers how such a system will be integrated into the relevant healthcare pathways and how expert users will interact with the system. Formal arguments are disconnected from the kinds of reasoning and arguments the average person engages in (whether that be an average expert user or average member of the public) as well as from the real world precisely because they are abstractions of both the reasoning process and the world. The result is that formal arguments are largely concerned with the *theory* of reasoning and arguments and less concerned with the actual *practice* of reasoning and arguing.

[7] Or more accurately we might call these the rules of inference.

2.3.2 Informal arguments

While there are slightly different positions that the proponent of informal arguments may take up, we submit that there are certain shared attitudes that are characteristic of the informal argument advocate. These are chiefly a "focus on the natural language arguments used in public discourse, clothed in their native ambiguity, vagueness and incompleteness," a "commitment to the study of argumentation as a dialectical process" and a "conviction that there are standards, norms, or advice for argument evaluation that is at once logical—not purely rhetorical or domain-specific—and at the same time not captured by the categories of deductive validity, soundness and inductive strength" [26]. So let us examine the advantages and disadvantages of informal arguments in turn and how they bear on supporting claims about safety.

2.3.2.1 Advantages and disadvantages Informal arguments, written as they are in natural language, are often accessible to a wide audience. Consider the differences between Gödel's proof of his incompleteness theorem and the argument that a clinician might give to convince their patient that an AI-enabled extubation system is safe to use. The former is nigh on incomprehensible to all but the most studied logicians and mathematicians whereas the latter is (or should be) readily comprehensible by the patient the clinician is treating. Moreover, because of their accessibility, informal arguments are often easier to both understand and evaluate. Commenting on or critiquing something like Gödel's proof is difficult not only because of the specialized knowledge of logic and mathematics required, but also because it is essentially written in an entirely different language. Informal arguments in contrast can engage a much wider audience because they are written out in natural language and tend to presume a general, not specialist, level of knowledge.

Another advantage of informal arguments is their close connection to reasoning in practice. By that we mean that informal arguments mirror the ways in which people reason and argue in their normal everyday lives. This is because the prevailing attitude amongst proponents of informal arguments, and informal logic in general, is that we must theorise about and understand actual (i.e., real-life, ordinary, everyday) arguments in their native habitat of public discourse and persuasion. The result has been the articulation of "methods of identifying, analysing and evaluating reasoning, which do not primarily rely on the instruments or nomenclature of formal logic" [26]. One such method articulated by Stephen Toulmin draws on judicial practices. For Toulmin, and for safety engineers inspired by his views, it is not enough to merely distinguish between premises and conclusions. When engaging in practical reasoning and argument there are "a good half-dozen functions to be performed by different sorts of proposition" some of which can be identified as "claims, data, warrants, modal qualifiers, conditions of rebuttal, statements about the applicability or inapplicability of warrants, and others" [43].

Informal arguments however also suffer from certain disadvantages, one of which concerns their evaluability. In contrast to formal arguments of which the evaluation is largely context-independent and field invariant, the evaluation of informal arguments requires an appreciation of the context and field in which the argument is presented. As Toulmin highlights, the "standards for judging the soundness, validity, cogency or strength of arguments are in practice field-dependent" [43]. To use Toulmin's terminology, the kind of data that one might produce to support a claim in one context, e.g., to a colleague working in a specialized discipline, may require no further justification or legitimization. Consider the following utterance. "The CNN is safe to deploy because its accuracy for extubation decisions was quite high at an AUC-ROC value of 0.94." When uttered from one developer to another, both of whom are working on the same project of producing an AI-enabled extubation-decision system, such a claim may require no further elaboration. In another context however, e.g., when the developer is speaking with a clinician who will be the expert user of the system, one might need to produce, in addition to the data and the claim, a warrant, i.e., an explicit proposition registering the legitimacy of the step from the data to the claim [43]. In our example, the developer may have to explicitly state what a CNN is (convolutional neural network), what an AUC-ROC (area under the receiving operating characteristic curve) value is, why it is a measure of system accuracy, and how that relates to the system's safe operation. In short, the evaluation of informal arguments becomes more difficult the larger the difference there is between the interlocutors' epistemic backgrounds.

A related disadvantage is that informal arguments cannot establish the certainty of their conclusions. Toulmin again nicely describes how, when judged against ideal "deductive" standards, informal arguments "are irreparably loose and lacking in rigour; the necessities and compulsions which they can claim—physical, moral and the rest—are never entirely compulsive or ineluctable in the way logical necessity can be; while their impossibilities are never as utterly adamantine as good, solid, logical impossibility" [43]. Informal arguments, tied as they are to reasoning in practice, are as often, if not more so, about establishing the likelihood of a conclusion as they are about establishing the necessity of a conclusion. This is often through looser inductive or abductive reasoning processes. Certain warrants may permit us to argue unequivocally to a conclusion, but this is an ideal exception, not the norm. More often than not, warrants entitle us to draw conclusions only tentatively subject to possible

exceptions or conditions [43]. Such is the case when arguing about the safety of an AI-enabled system.

2.3.3 Fallible safety claims

Safety claims are ultimately fallible knowledge claims given that their justification provided vis-à-vis the safety case is grounded in informal argumentation.[8] Indeed the disadvantages of informal arguments mentioned above turn out to be desirable features when arguing about the safety of a system and, as we shall see, communicating the content of a safety case to an epistemically diverse audience. Let us consider once again the AI-enabled extubation-decision system already introduced. The claim that this system is acceptably safe to deploy in a given context (i.e., within a certain healthcare pathway in a particular hospital) is one that can be undermined by defeaters or, in Toulmin's terminology, conditions of rebuttal [43]. It is vitally important that claims about a system's safety be subject to scrutiny and re-evaluation. The AI-enabled extubation system for example may only be acceptably safe to operate in a particular hospital by expert users that have received training on how to use the system. Taking the system to a different hospital or having untrained expert users utilise the system may undermine the claim that it is acceptably safe to operate.

Safety cases are also produced *for* someone, typically to persuade them to accept the claim that, for example, an AI-enabled extubation-decision system is acceptably safe to operate, on the basis of the arguments and evidence advanced in the safety case. As such they are part of a larger dialogue traditionally between developers, regulators and assessors. This dialectical spirit is another legacy of informal arguments that is desirable when discussing a system's safety. Though the evaluation of a safety case may be difficult for the layperson, designers and developers need not be restricted by the syntax and semantics of formal arguments when communicating safety claims which, ideally, should be closely examined, discussed and, if necessary, challenged. As mentioned above, designers and developers can, through the informal arguments in their safety case, tell a story about a system's safety, including what it means for the particular system to be safe, how it achieves this, and why one should be confident that the risks have been appropriately managed. For an AI-enabled extubation-decision system, the developers might emphasise to deployers (i.e., hospitals) and expert users (i.e., the clinicians working with the system) that the system is safe because it has reached a certain minimum threshold of accuracy, does not overwhelm the expert users with notifications and can be integrated into the existing healthcare pathway in such a way that it avoids unduly disrupting existing practices. In short, the developers will communicate their knowledge that the AI system is safe to the deployers and expert users, but more will be said about the form of this communication in Sect. 4.

3 AI system safety

While we have argued in the previous section that safety claims are descriptive fallible knowledge claims, one might object to this characterization. But this brings us to safety claims about AI-enabled systems in particular. Even if safety claims were not fallible knowledge claims, even if the aim of a safety case was to produce infallible knowledge, the nature of AI-enabled systems themselves precludes the possibility of obtaining infallible knowledge about their safety (or lack thereof). So in this section we look specifically at some of the features of AI-enabled systems that prevent one from making infallible knowledge claims about their safety. But first some terminological and clarificatory preliminaries.

Artificial intelligence (AI) is a widely used term with no clear boundaries, but it will suffice for our purposes to think of AI according to the definition given by the National Institute of Standards and Technology (NIST). AI, or an AI-enabled system, refers to the "capability of a device to perform functions that are normally associated with human intelligence such as reasoning, learning and self-improvement" [7].[9] There are roughly three components that together drive most current AI-enabled systems (including the AI-enabled extubation system that we have been using as an example throughout), and those are the deep neural network (DNN) (i.e., an artificial neural network with many layers of neurons between the input and output layers), the learning algorithm (which adjusts the weights between the neurons in the neural network) and the data (the largest portion of which serves as training data).

Additionally, it must be noted that AI is not a field, domain or industry but rather a *technology* that can be utilised in different fields, domains or industries. As such, AI is not, as one might be led to believe, some magical tool through which the world is objectively captured in a view from nowhere and revealed to us [1, 34]. On the contrary, AI-enabled systems do not necessarily learn anything "objective" about the world nor are they more "objective" in their decision making and behaviour than humans.

[8] Although it must be noted that there have been attempts to both formalise safety case arguments and incorporate features of formal arguments to complement features of informal reasoning employed in safety case arguments [15, 20, 38].

[9] Marvin Minsky similarly defined the study of artificial intelligence as "the science of making machines do things that would require intelligence if done by men [*sic*]" [33].

Moreover, AI-enabled systems, like any technology, are not created value free (i.e., normatively neutral). AI-enabled systems are conceived, designed and developed in a socio-political milieu and often by people or groups of people in positions of power [27]. Put simply, research in science and engineering is not value free, and this extends to the creation of AI-enabled systems. Though there is the potential for AI-enabled systems to benefit humanity we must be wary of the promises to this effect because they can just as easily perpetuate systemic biases and discriminate against those who are already marginalised and underrepresented [10, 11]. While it is important to recognise that there is a dominant narrative for AI that deserves scrutiny and criticism, such an analysis is outside the scope of this paper. So let us return to the peculiarities of AI-enabled systems. What exactly is it about AI-enabled systems that precludes the possibility of attaining infallible knowledge about their safety?

3.1 Uncertain behaviour, open contexts, black boxes and complicated consequences

It is impossible to know infallibly that an AI-enabled system is safe because of their uncertain behaviour, the varied open contexts in which they can operate and their opaque, often uninterpretable, inner workings. While the operating context can sometimes be spelled out in detail, this does relatively little to secure infallible knowledge about an AI system's safety. For example, one could be quite precise and detailed in outlining how an AI-enabled extubation system is supposed to be integrated into the relevant healthcare pathway, how expert users are supposed to interact with it and the limits of its capabilities. But while one can infer from this information that certain risks have been mitigated or addressed, it does little to justify infallible knowledge claims that a system will behave in a certain safe way, which is an issue that arises from the under-specificity of function of AI-enabled systems. Briefly, under-specificity of function refers to the gap that exists between the developer's intended goals for the system and the system's actual behaviour, sometimes known as "the semantic gap" [6]. This largely concerns the learning algorithm component of AI mentioned above. Such algorithms are often chosen not because they are the best or well understood, but because they work well enough.

So knowledge of the system's behaviour, let alone higher level properties like safety, is far from infallible, and this is only compounded by the opacity of AI-enabled systems. Transparency, as the term suggests, refers to the "visibility" of the system, in particular its inner workings, and the supposed logic or reasoning that the system employs to reach particular outputs. This largely concerns the DNN component of AI. While the inputs and outputs of the AI-enabled system are transparent, the same cannot be said for the many so-called hidden layers in the DNN. For example, though the expert user clinician might see that the AI-enabled extubation system takes as inputs features like the level of patient sedation and mode of ventilation and produces therefrom the output recommendation that the patient is ready for extubation, the clinician may have little understanding of *why* that particular output was produced and whether it is safe to proceed with extubation without any further investigation [25].[10] And even assuming that one has access to the DNN and can see all of the connections between the neurons and their weights, such transparency does not necessarily confer knowledge, let alone infallible knowledge, of the system's logic, even to those who designed it. Infallible knowledge about an AI-enabled system's safety is simply out of the question.

The different kinds of consequences arising from the use of AI-enabled systems also precludes the possibility of attaining infallible knowledge about their safety. On a strong interpretation, the consequences of using AI-enabled systems are different in kind, not merely by degree, from the consequences of using other technologies. This strong view is often adopted because AI-enabled systems, many argue, can lead via numerous paths to catastrophic or existential consequences [21]. Even on a weaker interpretation however, that the consequences of using AI-enabled systems are different merely in degree from other technologies, the consequences of utilising AI-enabled systems are such that they prevent one from obtaining infallible knowledge about their safety.

Claims about safety are, as we have seen, inherently context-dependent claims. But AI-enabled systems are increasingly general-purpose systems that are created with no specific use-context or operational environment in mind. The same AI-enabled system could be procured by many different deployers and adapted for different downstream uses [4]. The same AI-enabled system could just as easily be used to conduct post-cataract surgery follow-up calls with patients as it could conduct almost any other routine clinical conversation [29]. Similarly, the same AI-enabled system could just as easily generate predictions about when patients are ready for extubation as it could generate predictions about another critical and time-sensitive procedure. While it is in principle possible for some responsible party, e.g., the deployer, to outline the use-context or operational environment in detail and thereby identify and manage risks, this does little to justify an infallible knowledge claim that the system is acceptably safe. As mentioned above, these important contextual details at best only permit one to infer,

[10] Though it must be noted that there is a whole field of inquiry known as XAI (explainable AI) dedicated to investigating how AI-enabled systems can be rendered more transparent vis-à-vis explainability [2, 13, 25, 32].

in proportion to the detail given, that certain risks have been mitigated or addressed.

Even if the use-context or operational environment is specified in full detail, the sheer scale of the consequences of using AI-enabled systems precludes the possibility of attaining infallible knowledge about their safety. Consider that AI-enabled systems can have a large, even global, sphere of influence. AI-enabled systems can be copied and deployed *en masse* such that idiosyncrasies arising from design decisions made early in development influence the life chances and well-being of entire demographics. For example, an AI-enabled extubation system trained on data from patients seen at King's College Hospital in London, UK might be deployed for use in the University of Tokyo Hospital in Tokyo, Japan. In the best case scenario the AI-enabled extubation system is just as accurate in the former location as it is in the latter, but in the worse-case scenario patients are systematically misclassified in one or both locations as being ready for extubation when they are not because the training data was not representative of the demographics in the use-context.

So the sheer numbers of people that can be affected by the same AI-system makes even fallible knowledge claims about their safety difficult to establish. And the rising popularity and increased use of so-called "foundation models" only exacerbates this problem.[11] In short, foundation models are any models trained on a broad dataset that can then be adapted to perform a variety of downstream tasks. While not new per se, the scale and complexity of foundation models has increased to a point where they have begun to exhibit behaviours wholly unanticipated by their creators, an issue we have already touched on above [4]. More importantly, because foundation models require sophisticated hardware, immense processing power and gargantuan training datasets, this means that only a select few organisations are able to develop their own foundation models. This, coupled with the fact that many smaller organisations use these foundation models, albeit fine-tuning them for a particular task, means that there is a single point of failure for many different systems deeply rooted in the original foundation model. Once again, the nature of AI-enabled systems precludes the attainment of infallible knowledge about safety in practice.

The crux of the issue is that, for a number of reasons including some of which just discussed, it is impossible in practice to justify an infallible knowledge claim pertaining to the risk an AI-enabled system poses. Risk is often conceived of as the product of the likelihood and severity of a particular outcome. There is no reason why one could not, in principle, assess the likelihood and severity of an AI-enabled system's effects on a person's physical or psychological well-being for example. In practice however one is only more or less justified in fallibly knowing the risk of using an AI-enabled system given the inferences made from the available evidence to general claims about the likelihood and severity of specific outcomes obtaining. Note, however, that one might reason that we *can* possess infallible knowledge about the safety of AI-enabled systems. And indeed we can, but this is a triviality. If one assigns an all but certain likelihood or a high enough severity to the outcome, then it is trivial to say that one possesses infallible knowledge about the safety of an AI-enabled system, to wit, utilising the system will certainly lead to harm. Those who insist that we ought to worry about the existential consequences of AI-enabled systems fall into the latter camp, i.e., they assign an astronomical severity to the outcome of using AI-enabled systems. Their argument runs something like this. Even though the likelihood of developing paperclip-maximising superintelligent AI is infinitesimally small at the present, the severity of the consequences (namely human extinction) of doing so is such that we need to worry about preventing this outcome from obtaining *now*. Failure to address this problem through increased work on the value alignment problem, for example, will *certainly* lead to human extinction. So while it is possible to produce infallible knowledge claims about AI-enabled system safety, they are trivial claims that inherit their infallibility from some questionable premise.

4 Cooperative communication

Thus far we have argued that safety claims, when justified by a safety case, are descriptive fallible knowledge claims. This is in virtue of both the informal argumentation used in safety cases and the nature of AI-enabled systems themselves. Given that, the question with which we concern ourselves in this final section is how best to communicate this knowledge? The design and development of AI is increasingly participatory in nature as are assessments of their safety and ethical acceptability [5, 35]. This means that claims about AI-enabled system safety and their supporting arguments ought to be accessible to a wide range of affected stakeholders with different epistemic backgrounds. More specifically then, the question is how best to communicate this knowledge to an epistemically diverse group of stakeholders? How could one evaluate the calibre of the communication between different affected stakeholders? In what follows we suggest that the Gricean maxims of cooperative communication can be used to structure the form of the dialogue between stakeholders and also evaluate the calibre of the exchange.

[11] Explainer: What is a foundation model? https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

4.1 Inference and the cooperative principle

In his investigations of conversational implicature, Paul Grice famously outlined certain general features of discourse that are readily applicable to discourse of a specific kind, namely discourse surrounding the safety of AI-enabled systems [18]. Grice begins by noting that, for the kinds of conversations he is interested in analysing, the discourse is not random, i.e., communication does not consist of a series of disconnected remarks. There is, in general, some amount of cooperation between conversational partners given that each participant recognizes that in a given conversation, say a conversation about the safety of an AI-enabled system, there is, to some extent, "a common purpose or set of purposes, or at least a mutually accepted direction" in which the conversation moves [18]. There are therefore, at any given moment in a conversation, certain unsuitable "moves", i.e., conversational contributions. This leads Grice to formulate his Cooperative Principle: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" [18]. Though we shall refine this principle later given the specific direction of facilitating the effective exchange of knowledge about safety of an AI-enabled system, it is sufficient to note at the present that this Cooperative Principle can supply the essential structure for the exchange.

4.1.1 The Gricean maxims

Under the assumption of the Cooperative Principle, Grice draws out some additional maxims and submaxims that are worth describing given the role they will play in evaluating the calibre of the exchange. These maxims fall under four categories: quantity, quality, relation and manner. Under the category of quantity are the maxims to (1) communicate as much information as is required and (2) to refrain from disclosing any more information than is required [18]. Under the category of quality is the supermaxim to communicate truthfully, which can be achieved by adhering to the submaxims to (1) refrain from disclosing what one believes to be false and (2) refrain from disclosing what one lacks adequate evidence for [18]. Under the category of relation is the single and deceptively simple maxim, be relevant [18]. We say deceptively simple because, out of all of Grice's maxims, this is perhaps the most important for evaluating the calibre of the exchange of knowledge about an AI-enabled system's safety. More will be said about this below. Finally, under the category of manner is the supermaxim to be perspicuous which can be achieved by adhering to four submaxims: (1) avoid obscure expressions, (2) avoid ambiguity, (3) be brief and (4) be orderly [18].

4.1.2 Purposive communication

As alluded to above, the exchange that we are interested in analysing and evaluating is not random. The exchange between the communicator, e.g., the developer, and the communicatee, e.g., the expert user, is purposive. That is, the exchange is directed towards some end which, in our case, is to persuade the communicatee to believe in the justified claim (that we are assuming is true) that the AI-enabled system is safe to operate in a given context. The Gricean maxims are therefore ideal for thinking about the structure of the dialogue between communicator and communicatee given their intimate connection to purposive exchanges. Grice writes, "I have stated my maxims as if this purpose [of the conversation] were a maximally effective exchange of information," and while this is certainly not true of all conversations, it is undoubtedly the case when considering the disclosure of safety claims to affected stakeholders [18]. In short, communicators can ensure the effective exchange of safety knowledge claims by structuring their exchanges according to the Cooperative Principle which can be achieved via adherence to the Gricean maxims related to the categories of quantity, quality, relation and manner.

4.2 Evaluating knowledge exchanges

The Cooperative Principle and the Gricean maxims can provide the structure for the dialogue between communicator and communicatee, but the maxims in particular can also serve as criteria to evaluate the calibre of the knowledge exchange. While there is not much that can be said at an abstract or general level about the content of the knowledge exchange beyond what has already been mentioned, here we will draw on the example of the AI-enabled extubation-decision system to illustrate what an exchange between different affected stakeholders might look like. More importantly, we will highlight how the Gricean maxims can be used to evaluate the calibre of the exchange, particularly between stakeholders with different epistemic backgrounds.

4.2.1 Safety knowledge claims and AI-enabled extubation

Consider four different stakeholders: a developer, an expert user (i.e., a clinician), a prediction-recipient (i.e., a patient who will be affected by the use of the system by the clinician) and a representative of a regulatory agency. In the following imagined exchanges, one stakeholder will engage in a dialogue with another and attempt to persuade them to form the justified true belief that the AI-enabled extubation system is safe. So while the structure of the exchange will largely be consistent, i.e., adhering to the Cooperative Principle via the Gricean maxims, the content of the communication will differ given the different

epistemic backgrounds of the communicatees. The expert user for example may wish to know why the AI-enabled extubation system is safe and will not harm patients. We can imagine the following exchange (E1) between a clinician (the communicatee) and the developer (the communicator).

(E1) Clinician: *What features of the patient does the machine use to determine readiness for extubation?*

> Developer: *Primarily ones that align with clinical expectations, such as level of patient sedation and mode of ventilation.*

Note first the structure of this exchange. The quality and quantity of information disclosed by the developer are, at minimum, sufficient to facilitate the effective exchange of the information requested by the clinician. More might be said about the AI-enabled extubation system, i.e., the quantity of information could be increased (e.g., the developer could disclose that features like age, gender and ethnicity are features that the system *does not* use to determine patient readiness for extubation), but there is the risk that what is disclosed is increasingly *irrelevant* and would therefore degrade the calibre of the exchange. A similar risk exists if the quality of information is modified, i.e., what is disclosed is increasingly irrelevant. If, for example, the threshold that constitutes "truthful" communication is significantly changed then this might result in a different and less relevant exchange (e.g., the developer might add that the features that the system uses to determine readiness for extubation are in fact not locally important features specific to a particular decision, but globally important features given that they are averages derived from the whole training dataset) [25].

Contrast the exchange above between the clinician and developer with another example. Certain affected stakeholders may not have the opportunities to open a dialogue and speak with one another. Developers for example may not be able to communicate their knowledge of the system's safety to patients, and so communicating this knowledge may fall on the clinician as the expert user. Imagine the following exchange (E2) between a patient and/or their healthcare proxy (the communicatee) and their clinician (the communicator).

(E2) Patient: *How can you be sure that this system will correctly predict when I am ready for extubation?*

> Clinician: *While there are always risks with these types of procedures, this machine only assists me and makes predictions consistent with my clinical judgement. I am confident it will correctly predict when you will be ready for extubation.*

As with the first exchange, the form of this second exchange remains the same. That is, the quality, quantity, relevance and manner in which the information is disclosed by the communicator, in this case the clinician, are sufficient to facilitate the effective exchange of information requested by the communicatee, which in this case is the patient. Importantly, from the point of view of *evaluating* the calibre of this communicative exchange, what the clinician has disclosed in this exchange is *relevant* assuming the patient's non-medical and non-technical background. As mentioned above, relevance is crucial for ensuring optimal communication. If we assume that the patient has a different epistemic background, e.g., they are themselves a clinician or a software engineer working with artificial neural networks, then this imagined exchange may not satisfy the patient because it fails to be relevant. The justification provided by the clinician may fail in various ways to persuade the patient to believe in the truth of the safety claim. A patient with a background in medicine or computer science may, we might assume, be seeking a more technical response from the clinician. That is, they might be seeking an increased quantity of relevant information.

To see how exchanges like these might evolve into rich dialogues consider one more example. Imagine the following exchange (E3) between a representative of a regulatory agency and the developer.

(E3) Regulatory agent: *Why are patient features like ethnicity, gender and age included as inputs? Won't they bias the predictions your AI-enabled extubation system makes?*

> Developer: *These features were part of the training dataset and we simply left them as inputs. However our analysis of the system shows that these features have an importance near zero and so they likely have a negligible effect on the predictions.*

Again, the structure of this exchange is dictated first and foremost by relevance, i.e., what information is most relevant for the communicatee given their question but primarily epistemic background. The developer's answer is directly related to the regulatory agent's question. Similarly, the quality and quantity of information is such that an effective exchange of information is facilitated, and indeed may prompt follow up questions. The regulatory agent may, for example, be concerned about the possibility of systemic

discrimination by the AI-enabled extubation system and pursue this line of inquiry.

(E3*) Regulatory agent: *Is there any evidence that a machine formally "blind" to protected characteristics like ethnicity, gender and age performs differently?*

> Developer: *At present we have not trained a machine formally "blind" to these features, however we have reason to believe that excluding these features would not mitigate the risk of bias in the predictions.*

In addition to relevance, there is always a burden on communicators, the developer in this example, to communicate honestly, i.e., abide by the Gricean maxims that fall under the category of *quality*. In this follow up exchange (E3*) the developer is speaking truthfully by both admitting that they have not created a machine that excludes the inputs of ethnicity, gender and age, and by warning the regulatory agent that there is no guarantee that such a machine will be less biassed in its predictions. Indeed there is ample evidence, some of which may appear in a safety case prepared by the developer to justify claims about their AI-enabled extubation system's safety, to suggest that protected characteristics can be implicit in other unprotected characteristics and thereby render any exclusion of protected characteristics from the system's inputs meaningless at best [31].

Of course the quality of the communication must be balanced against the quantity, and that balance is evident in this and the above examples. While the developer could go on in conversation (E3*) about how and why they believe that excluding protected features would not mitigate the risk of bias and harm to certain groups of people, such additional information is not necessarily relevant given the context and would lower the calibre of the exchange were it to be included. Moreover, the regulatory agent could simply direct the exchange in that direction should they desire to understand more of the justification underpinning claims about how and why excluding protected characteristics might not mitigate the risk of bias. Indeed, communication about safety knowledge claims between different affected stakeholders ought to continuously occur in much the same way that the design, development and deployment of AI-enabled systems continuously occurs.

To sum up this section, the Gricean maxims can be used to evaluate the calibre of the exchange between different affected stakeholders when one is communicating knowledge about an AI-enabled system's safety to the other. High calibre exchanges are ones that, from the communicatee's point of view, are maximally efficient in persuading them to accept the justified true belief held by the communicator that the AI-enabled system is safe in whatever respect concerns the communicatee. From an outside perspective, we can evaluate the calibre of the exchange between communicator and communicatee using the Gricean maxims. High calibre exchanges are ones that are first and foremost relevant, i.e., the communicator tailors their communication to the particular communicatee they are engaging with given the communicatee's particular epistemic background. Second, high calibre exchanges appropriately balance the quantity and quality of information shared. Lastly, and this was largely implicit in the example exchanges given above (i.e., maxims connected to the category of manner were not violated), high calibre exchanges are communicated in an appropriate manner, i.e., the exchange is orderly, involves jargon that is appropriate and is just generally lucid.

5 Conclusion

AI-enabled systems are beginning to permeate our lives and society. They are used in virtually every sector and increasingly in safety–critical contexts [40, 44]. Assuring the safety of systems used in critical contexts is not a new activity. But what is new is the safety assurance of AI-enabled systems and, moreover, the communication of claims about the safety of AI-enabled systems to an epistemically diverse group of stakeholders. In this paper we have argued that safety claims, when justified by a safety case, are descriptive fallible knowledge claims. Even if the aim of a safety case was to justify infallible knowledge about the safety of a system, such infallible safety knowledge is impossible to attain in the case of AI-enabled systems. By their nature AI-enabled systems preclude the possibility of obtaining infallible knowledge concerning their safety or lack thereof. Finally, we have suggested that one can communicate knowledge of an AI-enabled system's safety by structuring their exchange according to Paul Grice's Cooperative Principle which can be achieved via adherence to the Gricean maxims of communication. Furthermore, these same maxims can be used to evaluate the calibre of the exchange, with the aim being to ensure that communicating knowledge about an AI-enabled system's safety is always of the highest calibre. In short, that the communication is relevant, of sufficient quantity and quality, and communicated perspicuously. Ultimately, the participatory nature of AI-enabled system design, development and assessment will require confronting the problem of how best to communicate safety claims to an epistemically diverse group of stakeholders. We hope that this paper represents one step towards addressing that problem.

Acknowledgement This work was supported by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York, and the UKRI project (EP/W011239/1) "Assuring Responsibility for Trustworthy Autonomous Systems".

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Adam, A.: Artificial Knowing: Gender and the Thinking Machine. Routledge (2006). https://doi.org/10.4324/9780203005057
2. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
3. Bloomfield, R., Netkachova, K., Stroud, R.: Security-informed safety: if it's not secure, it's not safe. In: Gorbenko, A., Romanovsky, A., Kharchenko, V. (eds.) Software engineering for resilient systems, vol. 8166, pp. 17–32. Springer, Berlin Heidelberg (2013). https://doi.org/10.1007/978-3-642-40894-6_2
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Liang, P.: On the opportunities and risks of foundation models (2022) (arXiv:2108.07258; Issue arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258
5. Burr, C., Leslie, D.: Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. AI Ethics **3**(1), 73–98 (2023). https://doi.org/10.1007/s43681-022-00178-0
6. Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., Porter, Z.: Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artif. Intell. **279**, 103201 (2020). https://doi.org/10.1016/j.artint.2019.103201
7. Computer Security Division, I.T.L.: CSRC Topic: Artificial intelligence | *CSRC*. CSRC | NIST. (2019) https://csrc.nist.gov/Topics/technologies/artificial-intelligence
8. Dekker, S.: Foundations of safety science: a century of understanding accidents and disasters. CRC Press, Taylor & Francis Group (2019)
9. Fernandes, M., Vieira, S.M., Leite, F., Palos, C., Finkelstein, S., Sousa, J.M.C.: Clinical decision support systems for triage in the emergency department using intelligent systems: a review. Artif. Intell. Med. **102**, 101762 (2020). https://doi.org/10.1016/j.artmed.2019.101762
10. Garvey, C.: Broken promises and empty threats: The evolution of AI in the USA, 1956–1996. Technol. Stories, 6(1) (2018)
11. Gebru, T.: Race and gender. In: Dubber, M.D., Pasquale, F., Das, S. (eds.) The oxford handbook of ethics of AI, pp. 251–269. Oxford University Press (2020). https://doi.org/10.1093/oxfordhb/9780190067397.013.16
12. Gettier, E.L.: Is justified true belief knowledge? Analysis **23**(6), 58–59 (1963). https://doi.org/10.1093/analys/23.6.121
13. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89 (2018) https://doi.org/10.1109/DSAA.2018.00018
14. Goldman, A.I.: Reliabilism and Contemporary Epistemology: Essays. Oxford University Press (2012)
15. Graydon, P.J.: Formal assurance arguments: a solution in search of a problem?. In 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, pp. 517–528. (2015) https://doi.org/10.1109/DSN.2015.28
16. Graydon, P.J.: The safety argumentation schools of thought. In AAA 2017 International Workshop on Argument for Agreement and Assurance, NF1676L-27810 (2017)
17. Greenwell, W.S., Knight, J.C., Holloway, C.M., Pease, J.J.: A taxonomy of fallacies in system safety arguments. In 24th International System Safety Conference (2006)
18. Grice, H. P.: Logic and conversation. In Cole, P., Morgan, J.L. (eds), Speech Acts, pp. 41–58. BRILL (1975) https://doi.org/10.1163/9789004368811_003
19. Habli, I.: On the Meaning of AI Safety [Working Paper]. ARRAY(0x556a925ef0f8). (2024) https://eprints.whiterose.ac.uk/204545/
20. Habli, I., Kelly, T.: Balancing the formal and informal in safety case arguments. VeriSure: Verification and Assurance Workshop, Colocated with Computer-Aided Verification (CAV) (2014)
21. Hendrycks, D., Mazeika, M., Woodside, T.: An Overview of Catastrophic AI Risks. (2023) https://doi.org/10.48550/ARXIV.2306.12001
22. Hollnagel, E., Wears, R. L., Braithwaite, J.: From Safety-I to Safety-II: a white paper (2015)
23. Hollnagel, E., Woods, D.D., Leveson, N.G.: Resilience Engineering: Concepts and Precepts. Ashgate (2006)
24. Hoyningen-Huene, P.: Context of discovery and context of justification. Stud. Hist. Philos. Sci. Part A **18**(4), 501–515 (1987). https://doi.org/10.1016/0039-3681(87)90005-7
25. Jia, Y., McDermid, J., Lawton, T., Habli, I.: The role of explainability in assuring safety of machine learning in healthcare. IEEE Trans. Emerg. Top. Comput. **10**(4), 1746–1760 (2022). https://doi.org/10.1109/TETC.2022.3171314
26. Johnson, R.H., Blair, J.A., Govier, T., Groarke, L., Hoaglund, J., Tindale, C.W.: The Rise of Informal Logic: Essays on Argumentation, Critical Thinking, Reasoning, and Politics. University of Windsor (2014)
27. Katz, Y.: Artificial Whiteness: Politics and Ideology in Artificial Intelligence. Columbia University Press (2020)
28. Kelly, T.P.: Arguing safety—a systematic approach to managing safety cases. PhD Thesis, Department of Computer Science, University of York (1998).
29. Khavandi, S., Lim, E., Higham, A., de Pennington, N., Bindra, M., Maling, S., Adams, M., Mole, G.: User-acceptability of an automated telephone call for post-operative follow-up after uncomplicated cataract surgery. Eye (2022). https://doi.org/10.1038/s41433-022-02289-8

30. Knight, J.C.: Safety critical systems: Challenges and directions. In Proceedings of the 24th International Conference on Software Engineering - ICSE '02, p. 547. (2002) https://doi.org/10.1145/581339.581406
31. Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., Yu, H.: Accountable algorithms. Univ. Pa. Law Rev. **165**(3), 3 (2017)
32. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
33. Minsky, M.: Semantic Information Processing. MIT Press. (1968) https://books.google.co.uk/books?id=F3NSAQAACAAJ
34. Nagel, T.: The View From Nowhere. Oxford University Press (1986)
35. Porter, Z., Habli, I., McDermid, J., Kaas, M.: A principles-based ethics assurance argument pattern for AI and autonomous systems. AI Ethics (2023). https://doi.org/10.1007/s43681-023-00297-2
36. Reed, B.: How to think about fallibilism. Philos. Stud. **107**(2), 143 (2002). https://doi.org/10.1023/A:1014759313260
37. Reed, B.: Certainty. In Zalta, E.N. (ed.), The Stanford Encyclopedia of Philosophy (Spring 2022). Metaphysics Research Lab, Stanford University. (2022) https://plato.stanford.edu/archives/spr2022/entries/certainty/
38. Rushby, J.: Formalism in safety cases. In: Dale, C., Anderson, T. (eds.) Making Systems Safer, pp. 3–17. Springer, London (2010). https://doi.org/10.1007/978-1-84996-086-1_1
39. Safety and functional safety. (2024). https://www.iec.ch/functional-safety
40. Savage, N.: The race to the top among the world's leaders in artificial intelligence. Nature **588**(7837), S102 (2020). https://doi.org/10.1038/d41586-020-03409-8
41. Sujan, M.A. et al.: Using safety cases in industry and healthcare—The Health Foundation (2012) https://www.health.org.uk/publications/using-safety-cases-in-industry-and-healthcare
42. Sujan, M.A., Habli, I., Kelly, T.P., Pozzi, S., Johnson, C.W.: Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices. Saf. Sci. **84**, 181–189 (2016). https://doi.org/10.1016/j.ssci.2015.12.021
43. Toulmin, S.: The Uses of Argument, Updated Cambridge University Press, Cambridge (2003)
44. United Nations Activities on Artificial Intelligence (AI). (2021).
45. Weisberg, E.M., Chu, L.C., Fishman, E.K.: The first use of artificial intelligence (AI) in the ER: triage not diagnosis. Emerg. Radiol. **27**(4), 4 (2020). https://doi.org/10.1007/s10140-020-01773-6