On the Meaning of AI Safety

Ibrahim Habli
Centre for Assuring Autonomy and UKRI AI Centre for Doctoral Training in Safe AI Systems (SAINTS)
University of York
York, United Kingdom
ibrahim.habli@york.ac.uk

***Abstract*—There is a growing urgency and global demand to address Artificial Intelligence (AI) safety. However, swift and sustained progress is unlikely to emerge without a shared understanding of what it means for the use of AI to be safe. This paper advances a comprehensive definition of AI safety and explores the fundamental concepts underpinning this definition. The aim is to contribute to a meaningful and inclusive discussion and further the public discourse on AI safety.**

Keywords—AI safety, autonomy, harm, risk

I. Introduction

Artificial Intelligence (AI) is here to stay. The technology now supports everyday activities, from routine tasks such as driving, to specialised decisions such as clinical diagnosis. The domains where the impact of AI could be most beneficial, including transport, health and social care, are safety critical. The recent advances in Large Language Models (LLMs) and foundational models have intensified the conversation around AI safety specifically [1], and responsible AI generally. These powerful models hold the promise of global benefits, but their uncontrolled use across society and the economy could have large-scale and catastrophic consequences.

A fundamental concern therefore arises: What does it mean for AI to be safe? The somewhat vague but commonly provided response is, '*it depends*', for example on where and how the technology is used and how it is developed. This paper proposes a well-rounded definition of AI safety. It then explores key concepts that influence its meaning. The aim is to inform the cross-disciplinary debate and advance the safety argument about AI.

II. Defining AI Safety

The definition of AI safety put forward in this paper is as follows:

Freedom** from **unacceptable risk** of **harm caused** by the **use** of **AI

Here, safety is characterised as a negative condition where freedom from harm is the focus. This definition is consistent with the assumption that rarely are complex technologies or interventions absolutely safe. As such, classic definitions of safety appeal to the *absence* of, or *freedom* from, unacceptable risk (see William Lowrance's seminal work on acceptable risk in 1976 [2]). Here, it is important to highlight both the objective and measurable facet of risk, e.g. as the product of likelihood and severity, and the subjective side of it, e.g. judging acceptability to different affected people.

In contrast, a more affirmative description of safety, emphasising the existence of protective capabilities, can be articulated as follows: *Protection from unacceptable risk of harm caused by the use of AI*. These definitions are interwoven. In the latter definition, the protective capability, often achieved through constant technological and social adjustments to changing and uncertain contexts, is intended to produce the *freedom* from unacceptable risk as outlined in the former definition (which is the focus of this paper).

III. Explaining AI Safety

Each key concept is next explained individually, acknowledging any interrelated aspects. For a visual summary of this discussion, please refer to Fig. 1.

A. Artificial Intelligence (AI)

AI, according to the National Institute of Standards and Technology, is defined as the "*capability of a device to perform functions that are normally associated with human intelligence such as reasoning, learning, and self-improvement*" [3]. The dominant AI technique driving most current AI-enabled capabilities is Deep Learning (DL). In its simplest form, DL is a neural network with multiple connected layers, trained on large datasets. DL serves as the core technology for developing LLMs. Two characteristics of AI, and notably DL and LLMs, present significant challenges to existing safety practices and standards: the under-specificity of the function and the opacity of the model.

Under-specificity refers to the gap between, on one hand, the underlying human intentions for deploying AI and, on the other, the specific, tangible specifications used to develop the technology [4]. Under-specificity hinders domain specialists, engineers and regulators in their efforts to establish and evaluate concrete safety requirements against which AI functions can be developed and tested. This challenge is exacerbated by the overwhelming focus in the literature on overall AI performance, overlooking nuances and context, e.g. treating historic, and out-of-context, clinician performance as a primary benchmark for clinical AI systems, which may not be appropriate for new or unforeseen situations [5].

The second challenge is opacity [6], and the lack of human-centred explanation [7]. The EU defines explainable AI as "*the ability of AI systems to provide clear and understandable explanations for their actions and decisions. Its central goal is to make the behaviour of these systems understandable to humans by elucidating the underlying mechanisms of their decision-making processes*" [8]. The inability to understand and explain how AI arrives at its outputs makes traceability and accountability challenging. It weakens our capacity to "*explain*" and "*deal with the consequences*" of AI functions [9].

In particular, LLMs are often presented as general-purpose AI models, which "*perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models"* [10]. However, this emphasis on generality in these 'Frontier AI' models, aiming for broad applicability and transferability, introduces inherent under-specificity. This, in turn, weakens our ability to proactively integrate safety constraints and generate concrete safety evidence before deployment, when it is most effective to assure safety.

979-8-3315-3741-8/25/$31.00 ©2025 IEEE
DOI 10.1109/EDCC-C66476.2025.00055

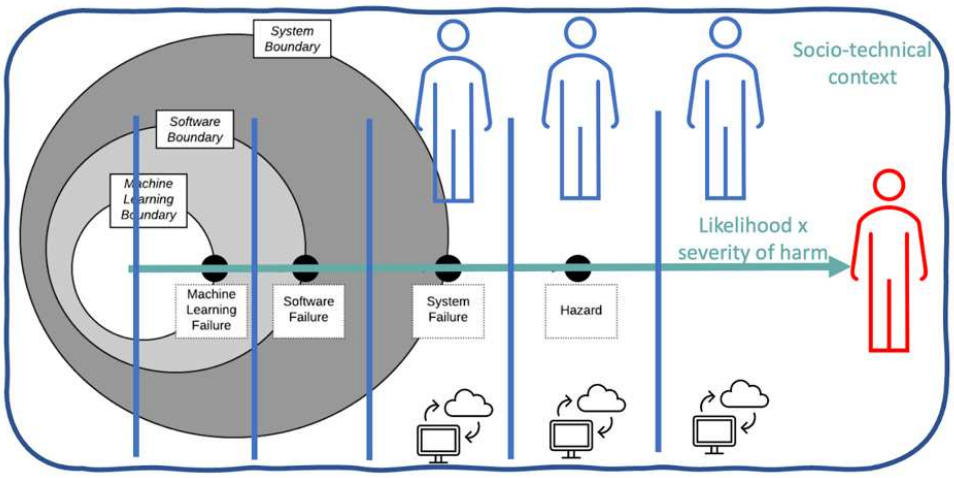

Fig. 1. A visual exploration of the AI definition, with a machine learning failure triggering a complex chain of events leading to harm (simplified here), countered by protective technical, human and social lines of defence

B. AI Context

The **use** of AI considers the algorithm in its intended technological, physical and social context. Safety, as a whole system property, is inherently sensitive to its context. The notion of AI context is varied, and covers how AI interacts with (1) other software components, e.g. cloud services, (2) hardware devices, e.g. CT scanners or lidars (3) the broader physical environment, e.g. communication between self-driving cars and the road infrastructure, (4) humans, e.g. capacity of doctors to detect bias in AI-based diagnosis systems and (5) the socio-political context in which AI is deployed, e.g. overreliance on technology to compensate for staff shortages. Despite the expectation that AI functions are adaptive, and deliver contextually meaningful experiences, the technology often exhibits brittleness [11].

AI is particularly susceptible to being '*fooled*' or '*confused*' by small and irrelevant environmental factors, such as stickers on stop signs [12]. Interestingly, while we often focus on external context, for AI, context extends not just outward but also inward. AI predictions and recommendations are inseparable from the contextual biases and stereotypical associations encoded in the training and testing datasets. This is noticeably the case for foundational models generated from unfiltered and massive datasets sourced from the Internet [13]. The multiple facets of this contextual complexity often weakens the robustness of AI safety evidence, especially when AI is used in constantly changing environments, e.g. in patient triage services. An in-depth understanding of AI and its context is a prerequisite for considering the subsequent safety concepts.

C. AI Causation

Causation should be interpreted in a broad socio-technical sense, considering the complex web of social and technological influences that AI produces. The impact can be direct, as seen in end-to-end machine learning for driverless cars, when AI functions autonomously control the vehicle sensors and actuators, or indirect, such as in AI-based clinical decision support systems, where clinicians are expected to make the final decisions. Determining causation also requires an understanding of the entire AI supply chain: causation can arise from upstream data collection practices to downstream user interactions and societal influences (Fig. 2). What is also key in establishing causation is modelling and understanding the wider software and systems architecture, including the level of redundancy, diversity and monitoring built into the overall design (Fig. 3). This is important in determining the extent to which AI failures could propagate into wider system failures.

The opacity of AI, and the interactive complexity within its wide context, especially for general purpose LLMs, make it difficult to model and trace exact causes and effects (e.g. tracing unsafe medication recommendations to unreliable and promotional marketing material used for training these foundational models). This, in turn, challenges our ability to proactively mitigate risk and reactively hold people accountable for actual harms caused by AI. Anticipating AI's potential consequences (forward-looking causation) and explaining how it arrived at those outcomes (backward-looking causation) is essential for a proactive, transparent and responsible AI safety culture.

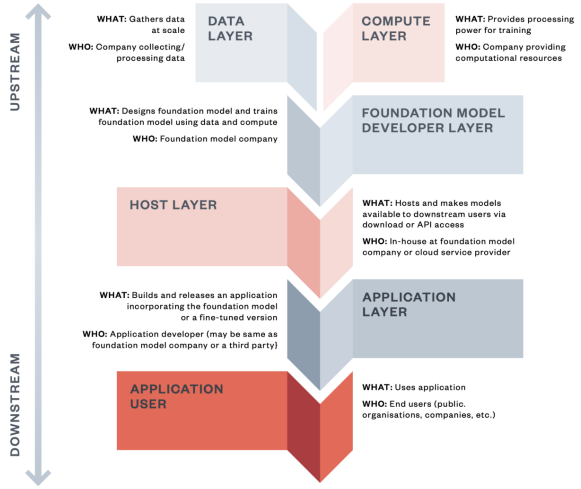

Fig. 2. A complex supply chain of foundational models [14]

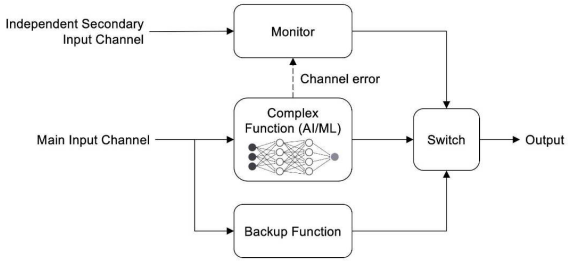

Fig. 3. AI system design, combined with backup/monitoring functions [15]

D. AI Harm

Harm in system safety is traditionally defined against physical damage. Typically, the focus is on damage to human physical health. This is followed by damage to property, with, more recently, the inclusion of damage to psychological health and to the environment. These remain key when considering AI safety. However, some discussion around AI safety seems to favour an '*expansive*' scope of harm [14], which stretches to discrimination, bias, misinformation, privacy violation and threats to democratic institutions, amongst other moral, political and social harms. These kinds of harms or wrongs are significant and concerning, regardless of whether they fall within the scope of AI safety or the broader scope of responsible or ethical AI. They should be systematically considered and, where causally relevant, integrated into AI safety assurance practices. For example, this includes avoiding safety measures that could unjustly limit personal freedom or exacerbate existing inequalities. Collaboration with the wider responsible or ethical AI community is essential to achieve this goal. Another important factor in safety assurance is intent: was harm intended, and if so, by whom? Was it justified? If harm is unintended, its occurrence is treated as a safety accident or incident. If harm is intended and this harm was caused maliciously, it is treated as a security event. Healthcare presents intriguing cases in this respect. Physical harm in surgery is often intended, for example making a precise incision, but may be justified, given the anticipated clinical benefit. In short, AI safety needs to build on, and where appropriate adapt, established methods from safety-critical domains. These specialist methods help us control both physical and psychological harm caused by AI.

E. AI Risk

Risk is the '*idea of a possibility of danger*' [16]. Technically, risk is the product of likelihood and severity of harm. However, risk is not an objective truth to be discovered and calculated. It is a social construct influenced by various uncertainties that are difficult to quantify [17], like the origin and quality of the data used to train AI or how users will actually interact with the tool. The notion of risk is central because complete avoidance of harm is rarely feasible. In risk analysis, harm is considered in relation to a particular context. Further, risk determination is typically framed by how harm could be caused "*in a stipulated way by the hazard*" [18], e.g. a hazard could be: misclassifying a 'slow down' traffic sign in foggy conditions. For narrow AI, i.e. intended to serve a specific purpose, hazard-based risk analysis is feasible though challenging. If AI's intended purpose is unclear or underspecified (e.g. classifying traffic signs in all weather conditions) and the AI model is opaque, it is hard to predict how likely it is to cause harm through its hazardous outputs. However, these concerns are significantly more complex for general-purpose AI, since the underpinning models, e.g. LLMs, are often presented as context-independent (i.e. specifying a well-defined purpose/context for this type of AI is often deliberately avoided by the AI developers). Even when context is identified for a specific use case, deployers of a general-purpose AI often lack sufficient access to the AI model and its vast training and testing datasets to allow them to accurately assess the likelihood of harm.

F. Risk acceptability

Acceptable risk *to whom* and *given what else* are two factors that need to be assessed as fundamental inputs into the AI risk decision-making process. Risk acceptability, and the lack of it, is a complex social notion not a technical one. To this end, risk decision-making needs to be participatory and transparent. Affected stakeholders, or their trusted representatives, e.g. regulators, need to be meaningfully involved in how the use of AI could present them, and others in society, with potential benefits and risks. The variety of risk communicated should be comprehensive, covering physical, psychological and societal ones, amongst others, to allow the affected stakeholders to understand and consider any necessary tradeoffs. This will enable an open and reflective dialogue about the distribution of benefits and risks from the use of an AI system and whether it is equitable across all affected stakeholders [19].

G. Confidence and communication

Freedom from unacceptable risk is rarely, if ever, a certainty. Rather, it is communicated with a degree of confidence. Confidence is determined given the effectiveness of the protection or control measures deployed, acknowledged uncertainties and underlying assumptions. For AI, *epistemic* uncertainty is particularly significant. It represents deficits in our *knowledge* about the AI implementation and outputs, and the impact the technology may have on its environment [20]. For instance, poor performance of an AI model in accurately diagnosing rare diseases often reflects our incomplete knowledge about these diseases and our limited exposure to these clinical scenarios. In safety, confidence may be effectively communicated using safety cases [21]. The explicit and structured arguments in safety cases provide a means for justifying and evaluating confidence about the absence of unacceptable risk. An AI safety case can help facilitate the scrutiny of the otherwise implicit reasoning, the interrogation of sufficiency of the evidence, and whether assumptions hold true (for whom and under what conditions). This, in turn, helps foster transparency throughout the entire AI lifecycle.

For example, in Fig. 4, we depict an ethics assurance argument, in which AI safety, as well as other principles such as equity and respect for human autonomy are considered. The argument, represented using the Goal Structuring Notation and explained in detail in [19], advances the claim (JG1) that the '*distribution of benefit, tolerable residual risk, and tolerable constraint on human autonomy (from use of AI) is equitable across all affected stakeholders*'. There are three key issues to note here: (1) The question of safety is hard, and counterproductive, to consider in isolation from the wider issue of fairness and equity (*JG1*). (2) Affected stakeholders should be identified (*JC4*), and their diversity considered, including the different kinds of (positive and negative) impact AI will have on their lives. (3) Tradeoffs are inevitable and should be reasoned about in a participatory manner (*JG5*).

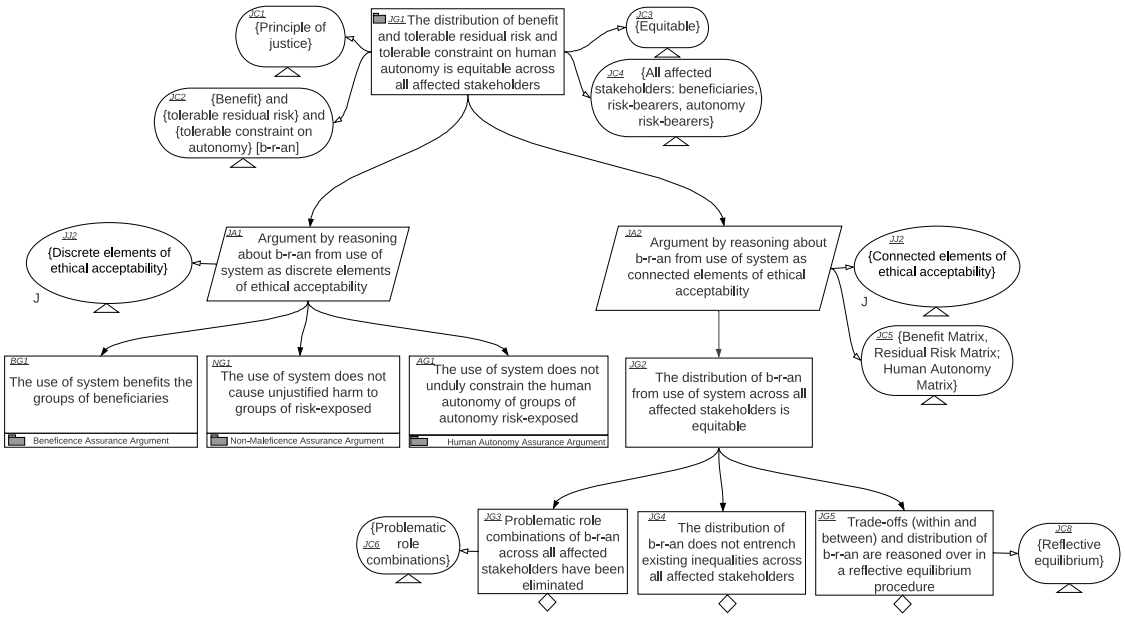

Fig. 4. Arguing safety in the context of a wider AI ethical assurance framework, depicted as a GSN argument pattern [19]

IV. Basic Ingredients for Authentic AI Safety

Just as a surgical checklist is not a complete guide for training competent surgeons, the definition of AI safety above is not an exhaustive tutorial on a rapidly emerging field. Its aim is to ensure that established safety concepts do not get lost in the hype surrounding AI, a field dominated by both a deliberate downplaying of real and pressing safety concerns and an unhealthy fixation on existential threats [22]. These core concepts are essential ingredients for building a responsible safety mindset, replacing the current sci-fi hubris with a pluralistic basis that upholds an equitable right to safety for all.

Acknowledgment

This work was supported by the UKRI AI Centre for Doctoral Training in Safe AI Systems (SAINTS) (EP/Y030540/1), UKRI project "Assuring Responsibility for Trustworthy Autonomous Systems" (EP/W011239/1) and the Centre for Assuring Autonomy, a partnership between Lloyd's Register Foundation and the University of York. Special thanks to Ana MacIntosh, Rob Alexander and Drew Rae for their valuable feedback.

References

[1] Y. Bengio, et al., "International Scientific Report on the Safety of Advanced AI," Department for Science, Innovation and Technology, 2024.

[2] W. Lowrance, "Of acceptable risk: Science and the determination of safety", 1976.

[3] "CSRC Topic: artificial intelligence | CSRC," *CSRC | NIST*, Oct. 28, 2019. https://csrc.nist.gov/Topics/Technologies/artificial-intelligence

[4] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, "Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective," *Artificial Intelligence*, vol. 279, p. 103201, Feb, 2020.

[5] E. J. Topol, "High-performance medicine: the Convergence of Human and Artificial Intelligence," *Nature Medicine*, vol. 25, no. 1, Jan. 2019.

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, May 2019.

[7] Y. Jia, J. McDermid, N. Hughes, M. Sujan, T. Lawton and I. Habli, "The Need for the Human-Centred Explanation for ML-based Clinical Decision Support Systems," *IEEE 11th International Conference on Healthcare Informatics (ICHI)*, Houston, TX, USA, 2023, pp. 446-452.

[8] "Explainable Artificial Intelligence," 2023. doi: 10.2804/132319

[9] Z. Porter, A. Zimmermann, P. Morgan, J. McDermid, T. Lawton, and I. Habli, "Distinguishing two features of accountability for AI technologies," Nature Machine Intelligence, vol. 4, no. 9, Sep. 2022

[10] "AI Safety Summit: introduction (HTML)," GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html

[11] L. Eliot, "Exposing The Brittleness Of Generative AI As Exemplified By The Recent Gibberish Meltdown Of ChatGPT," Forbes. Feb 25, 2024,

[12] D. Heaven, "Why deep-learning AIs are so easy to fool," Nature, vol. 574, no. 7777, pp. 163–166, Oct. 2019.

[13] E. Bender, A. McMillan-Major, S. Shmitchell, and T. Gebru, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ," FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Mar. 2021.

[14] M. Davies and M. Birtwistle, "Regulating AI in the UK," www.adalovelaceinstitute.org, Jul. 18, 2023. https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/

[15] J. Fenn, M. Nicholson, G. Pai, and M. Wilkinson, "Architecting Safer Autonomous Aviation Systems," *arXiv.org*, 2023. https://arxiv.org/abs/2301.08138 (accessed Oct. 28, 2024).

[16] R v Board of Trustees of the Science Museum [1993] 1 WLR 1171

[17] S. O. Hansson, "Risk: objective or subjective, facts or values," *Journal of Risk Research*, vol. 13, no. 2, pp. 231–238, Mar. 2010

[18] "Risk management: Expert guidance - Reducing risks, protecting people - R2P2," https://www.hse.gov.uk/enforce/expert/r2p2.htm

[19] Z. Porter, Ibrahim Habli, J. McDermid, and Marten, "A principles-based ethics assurance argument pattern for AI and autonomous systems," AI and ethics, Jun. 2023.

[20] J. Jiang, S. Kahai, and M. Yang, "Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty," *International Journal of Human-Computer Studies*, vol. 165, p. 102839, Sep. 2022.

[21] M. A. Sujan, I. Habli, T. P. Kelly, S. Pozzi, and C. W. Johnson, "Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices," *Safety Science*, vol. 84, Apr. 2016.

[22] J. Stilgoe, "Technological risks are not the end of the world," *Science*, vol. 384, no. 6693, Apr. 2024.