On the Meaning of AI Safety

Ibrahim Habli
UKRI Centre for Doctoral Training in AI Safety (SAINTS)
Centre for Assuring Autonomy
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. But 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, followed by an exploration of the fundamental concepts that underpin this definition. The aim is to contribute to a meaningful and inclusive discussion and further the public discourse on AI safety.

Keywords: Artificial Intelligence · Safety · Risk · Harm

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. In these domains, AI failures can have significant consequences, causing physical, psychological, and social harm. 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[1] [1].

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. In contrast, a more constructive and affirmative description, 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.

[1] Which could reveal that AI may be unsafe to deploy in particular contexts and why.

Explaining AI Safety

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

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*" [2]. 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. Two characteristics of AI, and specifically DL, present significant challenges to existing safety practices: 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 [3]. 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 [4]. The second challenge is opacity [5]. The inability to understand 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 [6]. Under-specificity and opacity pose challenges to assuring the safety of both narrow[2] and general-purpose[3] AI models.

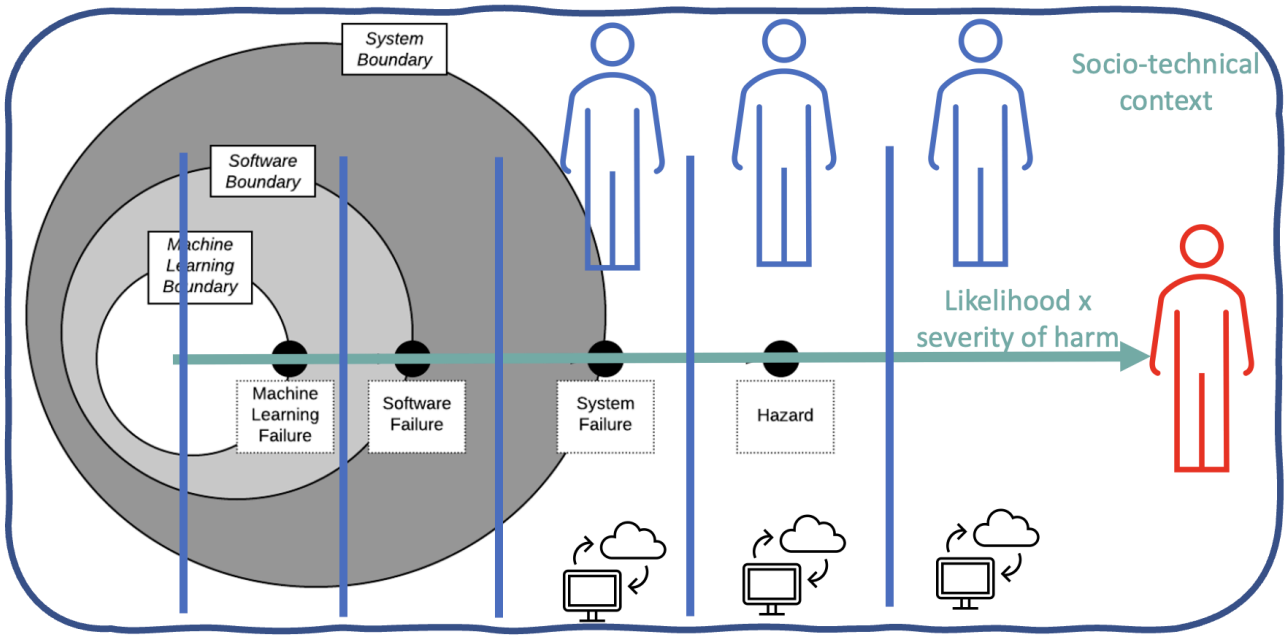

Figure 1: A visual exploration of the AI definition, with a machine learning failure triggering a complex chain of events leading to harm (simplified here), countered by protective technical, human and social lines of defence

[2] "*Narrow artificial intelligence (narrow AI) is AI that is designed to perform a specific task. It is a specific type of artificial intelligence in which a learning algorithm is designed to perform a single task or narrow set of tasks, and any knowledge gained from performing the task will not automatically be applicable or transferable*" [7].

[3] Also known as ' Frontier AI': "*It refers to highly capable general-purpose AI models, most often foundation models, that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models. It can then enable narrow use cases*" [7].

Use 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 encompasses how AI interacts with (1) other software components, e.g. cloud computing 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 [8]. AI is particularly susceptible to being '*fooled*' or '*confused*' by small and irrelevant environmental factors, such as stickers on stop signs [9]. 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 test data [10]. The robustness of AI safety evidence is challenged by the multiple facets of this contextual complexity. This often weakens the validity of such evidence in constantly changing environments, e.g. in urban driving or patient-centric healthcare delivery. An in-depth understanding of AI and its context is a prerequisite for considering the subsequent safety concepts.

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. Causation also requires an understanding of the entire AI supply chain: causation can stem from upstream data collection practices to downstream user interactions and societal influences. The opacity of AI, and the interactive complexity within its wide context, make it difficult to model and trace exact causes and effects. 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 actually arrived at those outcomes (backward-looking causation) is essential for an proactive, transparent and accountable AI safety culture.

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 well-being and to the environment. These remain key when considering AI safety. However, the discussion around AI safety seems to favour an '*expansive*' scope of harm [11] which stretches to discrimination, bias, misinformation, privacy violation and threats to democratic institutions, amongst other moral, political, social and financial harms. These kinds of harm are significant and concerning. They should be systematically addressed and mitigated in an integrated manner (e.g. avoiding safety measures that unjustifiably constrain personal freedom or entrench existing inequalities). Another important factor in system safety 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. As such, AI safety needs to be both expansive and specific. On the one hand, it should consider diverse kinds of harm, from social bias to physical accidents. This makes it an inclusive concept that everyone can contribute to. On the other hand, AI safety also needs to build on established methods from safety-critical domains. These specialist methods help us control both physical and psychological harm caused by AI.

Risk is the '*idea of a possibility of danger*' [12]. Technically, risk is the product of likelihood and severity of harm. However, risk is not an objective truth to be discovered and calculated. It's a social construct influenced by various uncertainties that are difficult to quantify, like the origin and quality of the data used to train AI or how users might 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*" [13], e.g. a hazard could be: misclassifying a 'slow down' traffic sign in foggy conditions. For narrow AI, 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 technology is often presented as context-independent (i.e. specifying a well-specified 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 access to the AI model and its vast training and testing datasets to allow them to accurately assess the likelihood of harm.

Unacceptable 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, moral and legal 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 [14].

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. In safety, confidence may be effectively communicated using safety cases [15]. 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.

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. 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.

Acknowledgements

This work was supported by UKRI AI Centre for Doctoral Training in Lifelong Safety Assurance of AI-enabled Autonomous Systems (SAINTS) (EP/Y030540/1), the 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. A year of racing ahead with AI and not breaking things. Nat Mach Intell 5, 1337 (2023). https://doi.org/10.1038/s42256-023-00782-7
2. National Institute of Standards and Technology (NIST), Definition of AI, accessed: 21 October 2023 https://csrc.nist.gov/Topics/technologies/artificial-intelligence
3. Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artificial Intelligence, 279, 103201.
4. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.
5. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.
6. Porter, Z., Zimmermann, A., Morgan, P., McDermid, J., Lawton, T., & Habli, I. (2022). Distinguishing two features of accountability for AI technologies. Nature Machine Intelligence, 4(9), 734-736.
7. Department for Science, Innovation & TechnologyAI Safety Summit, accessed: 21 October 2023, https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html
8. Cummings, M. L. (2020). The surprising brittleness of AI. Women Corporate Directors, 3.
9. Heaven, D. (2019). Why deep-learning AIs are so easy to fool. Nature, 574(7777), 163-166.
10. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).
11. Ada Lovelace Institute, Regulating AI in the UK (2023), accessed: 21 October 2023, https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/
12. R v Board of Trustees of the Science Museum [1993] 1 WLR 1171
13. Health and Safety Executive, Reducing risks, protecting people - R2P2, accessed 21 October 2023, https://www.hse.gov.uk/enforce/expert/r2p2.htm
14. Porter, Z., Habli, I., McDermid, J., & Kaas, M. (2023). A principles-based ethics assurance argument pattern for AI and autonomous systems. AI and Ethics, 1-24.
15. Sujan, M. A., Habli, I., Kelly, T. P., Pozzi, S., & Johnson, C. W. (2016). Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices. Safety science, 84, 181-189.