On the Meaning of AI Safety

Ibrahim Habli
Department of Computer Science
University of York
York, United Kingdom
ibrahim.habli@york.ac.uk

Abstract

In this paper, I propose a definition of AI safety. I then explore the fundamental concepts that underlie this definition. My aim is to contribute to a constructive discussion and further the discourse on AI safety.

Keywords: Artificial Intelligence · Safety · Risk · Harm

Introduction

AI systems now assist with everyday tasks, spanning from routine activities like driving to specialised decisions, such as clinical diagnosis. Yet, a fundamental concern arises: Is AI safe? The somewhat vague but commonly provided response is, 'it depends'[1]. Here, I propose a definition of AI safety. Next, I explore key concepts on which the meaning of AI safety depends. My aim is to help inform the debate and advance the safety argument about AI[2].

Defining AI Safety

My proposed definition of AI safety is as follows:

Absence *of* ***unacceptable risk*** *of* ***harm caused*** *by the* ***use*** *of* ***AI***

In this definition, safety is characterised as a negative condition where the absence of risk is the focus. Alternatively, 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 two definitions are interconnected. The protective capability in the latter definition is designed to result in the absence of risk as described in the former definition.

Explaining AI Safety

I will now address each key concept individually. I will also acknowledge any interrelated aspects of these concepts as necessary. For a visual summary of this discussion, please refer to Figure 1.

[1] Unfortunately, it is a common response by professionals to many complex questions!

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

Artificial Intelligence (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*" [1]. 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 layers, trained on large datasets. However, two characteristics of AI, particularly DL, present significant challenges to existing safety practices: the under-specificity of function[3] and the opacity of the model. Under-specificity represents the delta between the underlying human intentions for the deployment of AI and the specific, tangible specifications employed in constructing the technology. Under-specificity hinders engineers in their efforts to establish and evaluate concrete safety requirements against which AI functions can be developed and tested. This challenge is compounded by the overwhelming focus in the literature on overall AI performance. Opacity, i.e. the inability to understand how AI arrives at its outputs, complicates accountability, especially in terms of "*explaining*" and "*dealing with the consequences*" of AI functions and their development [3]. Under-specificity and opacity pose challenges to ensuring the safety of both narrow[4] and general-purpose[5] AI models.

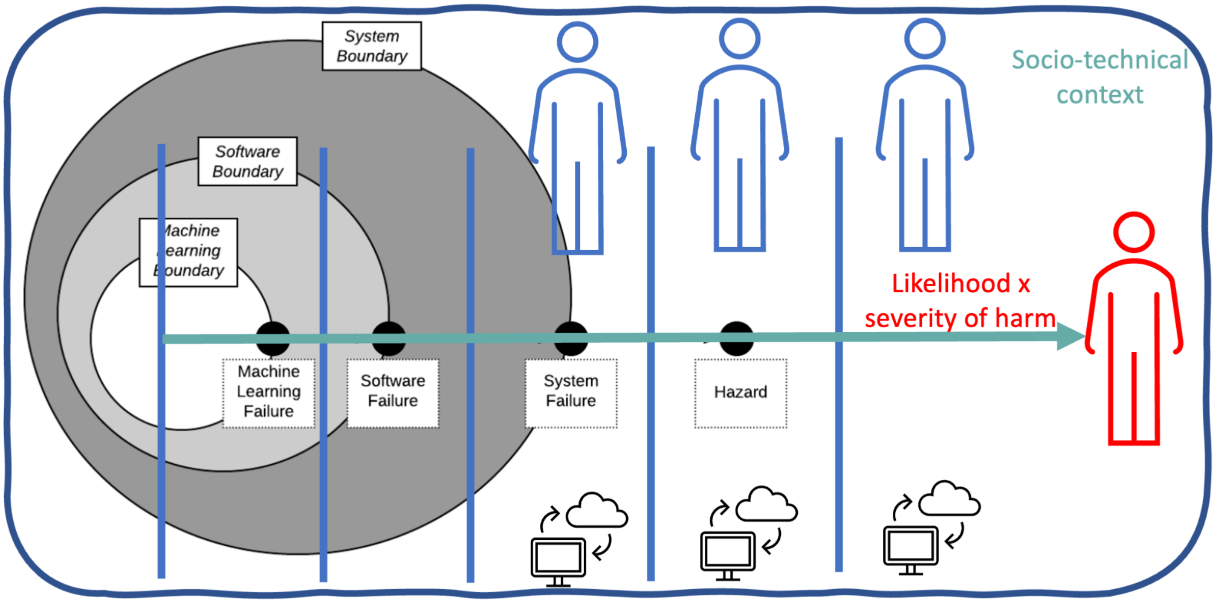

Figure 1: An Illustrative depiction of the AI definition, with a machine learning failure as a source of a hazardous chain of complex chain of events leading to harm (though here graphically simplified), with protective human and technical lines of defence that are used for risk reduction.

Use considers the algorithm or model in its intended physical and social context. Safety is context-sensitive. AI systems often exhibit brittleness and susceptibility to being '*fooled*' or '*confused*' by irrelevant environmental factors, such as stickers on stop signs [10]. The notion of context is multifaceted and varied, and includes how an AI component interacts with (1) other software components, e.g. cloud computing, (2) hardware components, e.g. image scanner devices, (3) the broader physical environment, e.g. communication between self-driving cars in vehicle platooning, (4) humans, e.g. how AI output is

[3] Also see the discussion of 'semantic gap' by Burton et al. [2].

[4] "*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*" [4].

[5] 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*" [4].

presented and 'explained' and (5) the social context in which AI is developed and deployed, e.g. staffing levels on clinical wards. The AI context is often described as "open" and "adaptive," as seen in scenarios like urban driving or patient-facing services. This introduces challenges in terms of the scale, quality and quality of the safety evidence required and the assumptions underlying its ongoing validity in a changing environment. Describing AI and its context is a prerequisite for considering the subsequent concepts.

Causation should be interpreted in its wider socio-technical sense, with the impact of AI on its physical and social context represented using complex cause-and-effect chains. The impact can be both direct, as seen in end-to-end machine learning for autonomous driving, and indirect, as in clinical decision support systems where clinicians make the final decisions and take actions. Causation should also cover the AI supply chains, upstream (what and how data is collected[6]), all the way downstream (how, and under what conditions, end-users and other people interact with, and are influenced by, the use of AI). The opacity of AI and the interactive complexity within its context complicate the modelling of causation. This in turn challenges our ability to proactively mitigate risk and reactively hold people accountable for harms caused by AI.

Harm in system safety is typically defined against physical damage. Traditionally, the focus is on damage to human physical health. This is followed by damage to property, with, more recently, the inclusion of damage to human psychological health and to the environment. A key perspective here 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 it involves malice, it is flagged as a security event. Healthcare and defence present complex cases in this respect. Harm there may be intended, say for surgery, but may be justified, given anticipated clinical benefit. In the AI literature, the discussion seems to favour an '*expansive*' scope of harm [5]. The particular focus appears to be on 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 proactively addressed and mitigated in an integrated manner (e.g. avoiding safety measures that unjustifiably constrain personal freedom or entrench existing inequalities). However, we need to strike the right balance between AI safety, as an umbrella and loose term, encompassing all kinds of harm, and AI safety as a specialist term, with its links to the system safety field and its well-established methods and particular focus on physical and psychological harm.

Risk is the '*idea of a possibility of danger*' [6]. More commonly, risk is the product of likelihood and severity of harm. The term risk is central here because complete avoidance of harm is rarely feasible. In risk analysis, harm is considered in relation to a particular context. Further, harm is often framed in relation to the use of a system and, more specifically, how system outputs may become sources of harm, i.e. hazards[7]. For narrow AI, hazard-based risk analysis is feasible though challenging. Underspecification in the definition of the intended output, combined with model opacity, makes it hard to estimate the likelihood of AI leading to harm via its potential hazardous output. For general-purpose AI, identifying harm, and its severity and likelihood, may be infeasible since the technology is often presented as context-independent. 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 training and testing datasets to allow them to assess the likelihood of harm (due to potential hazardous outputs of the system).

[6] There is an implicit assumption in AI development that large datasets replace the need for detailed requirements. This is a root cause of under-specificity.

[7] In other words, risk determination is typically framed by how harm could be caused "*in a stipulated way by the hazard*" [7].

Unacceptable risk to whom and given what else are two factors that need to be assessed as foundational inputs into the AI risk decision-making process. Risk acceptability, and the lack of it, is a complex social construct not a technical one. As such, risk decision-making needs to be participatory. Affected stakeholders, or their trusted representatives, e.g. regulators, need to be meaningfully involved in how the use of AI could present them with potential benefits and risks. The risks communicated should be expansive, covering physical, psychological, moral and legal ones, amongst others, to allow the affected stakeholders to understand 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 affected stakeholders [8].

Absence of unacceptable risk is rarely, if ever, an absolute goal. Rather, it is communicated with a degree of confidence. Confidence is determined given the effectiveness of the **protection**, control or management measures deployed, evidence available, sources of uncertainty declared and assumptions made. For AI, under-specificity and opacity are significant sources of *epistemic* uncertainty. It represents deficits in our *knowledge* about the AI implementation and outputs and the impact AI may have on its environment. In safety, confidence may be communicated using safety cases. Such arguments provide a means for justifying and evaluating confidence in the safety of complex systems. An AI safety case should help facilitate the scrutiny of the otherwise implicit reasoning, the interrogation of sufficiency of the evidence, and the validity of the assumptions. This, in turn, helps foster increased transparency throughout the entire AI lifecycle [9][11].

Acknowledgements

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

References

1. National Institute of Standards and Technology (NIST), Definition of AI, accessed: 21 October 2023 https://csrc.nist.gov/Topics/technologies/artificial-intelligence
2. 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.
3. 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.
4. 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
5. Ada Lovelace Institute, Regulating AI in the UK (2023), accessed: 21 October 2023, https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/
6. R v Board of Trustees of the Science Museum [1993] 1 WLR 1171
7. Health and Safety Executive, Reducing risks, protecting people - R2P2, accessed 21 October 2023, https://www.hse.gov.uk/enforce/expert/r2p2.htm
8. 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.
9. The Health Foundation, Using safety cases in industry and healthcare, December 2012.
10. Heaven, D. (2019). Why deep-learning AIs are so easy to fool. Nature, 574(7777), 163-166.
11. Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., & Habli, I. (2021). Guidance on the assurance of machine learning in autonomous systems (AMLAS). arXiv preprint arXiv:2102.01564.