AI Safety: Navigating the Expanding Landscape of Potential Harms

Ibrahim Habli and John McDermid look at how the emergence of Artificial Intelligence (AI) is adding pressure on the safety community to rethink the boundaries of safety engineering. The rise of AI is exposing a wider spectrum of harms, with significant moral, social and economic impacts. Can the safety-critical community realistically expand the scope of safety to address these intangible yet impactful AI-induced harms?

Traditional safety management focuses on preventing accidental harm. But given that absolute prevention is rarely possible, the focus of safety activities tends to shift to risk, i.e. the likelihood and severity of harm. The kinds of harm of particular concern are typically physical in nature, directed to humans, as well as to property and the environment. The increasing use of Artificial Intelligence (AI), whether in safety-critical applications or not, extends the scope further. It raises concerns about the potential of the technology to cause other significant harms [1]: moral, social, political, and economic [2]. This has led to growing pressure and expectations for "safety" measures to address these wider categories of harm [3], and the term "AI Safety" is often used to encompass managing the risks from AI for these wider categories of harm. Whilst such concerns might seem to fall naturally within the remit of safety engineering, can the safety-critical community realistically expand the scope of safety to address these intangible yet impactful AI-induced harms? Is such an expansive approach feasible, and if so, what strategies will be necessary to navigate this vast and uncharted territory?

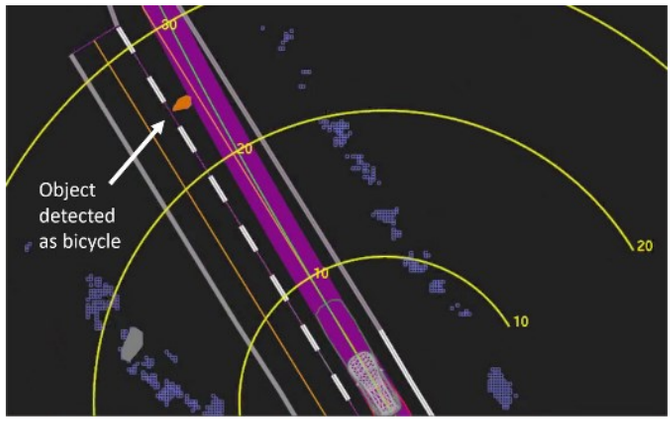
From Physical to Psychological and Environmental Harm

Safety-critical industries such as transport, defence and manufacturing have traditionally placed unique emphasis on eliminating or reducing the risk of physical harm caused by sociotechnical systems on humans and on property. Death, injury, and damage to physical artefacts have been the core categories of concern, supported by a rigorous ecosystem of standards, guidelines, and best practices, which have gradually been evolved by the relevant subject-matter and professional communities.

Further, in the last two decades, we have witnessed a critical and growing recognition of the psychological dimensions of harm. In particular, the development of highly interactive digital technologies and software-intensive automation systems, while enhancing performance and physical safety, has introduced or shed new lights on psychological risk factors. The potential for anxiety, stress and trauma, particularly in high-pressure environments and complex human-machine interactions, is now better understood and proactively analysed [4]. Examples include air traffic controllers or clinicians experiencing fatigue and burnout due to complex workflows or understaffed settings.

This historically narrow scoping of harm has undergone, albeit belatedly, an expansion in recent years, leading to the inclusion of environmental impacts within the scope of safety management [5]. Driven by the climate crisis, issues such as ecological damage and pollution are now better recognised as potential consequences of activity in safety-critical industries, particularly in the transport and energy sectors.

AI systems, mostly operating in 'testing' phases or 'advisory' roles, have been linked to actual physical, psychological and environmental harms. For example, in 2018, an Uber test vehicle, operating in a self-driving mode, hit and killed Elaine Herzberg in Tempe, Arizona [6]. This fatal accident illustrates how AI-enabled features could contribute, amongst other key factors, to human harm. The Automated Driving System (ADS), which relied on machine learning, inaccurately classified Herzberg and her trajectory.



Automated crash avoidance, a function of the ADS, was not possible by the time the system predicted that a crash was imminent. While the deficiencies in the design and risk assessment of ADS were highlighted as a contributory factor, the National Transportation Safety Board (NTSB) investigation identified the *'failure of the vehicle operator to monitor the driving environment and the operation of the automated driving system'* as the '*probable cause of the crash'.* The complexity of the human-machine interaction and Uber's safety culture, increased the risk of 'automation complacency', whereby the vehicle operator over-relied on the capability of ADS.

Intent Matters

In safety management, the focus is primarily on *accidental* harm, arising from errors or mistakes. Any harmful consequences are unintended, in contrast to *deliberate* harm caused by malicious actors. This has formed a basis for the conceptual difference between safety (freedom from accidental harm) and security (protection from malicious harm). It has led to

different specialised methods such as hazard analysis in safety engineering and attack trees in security engineering. However, recognising this difference alone is counterproductive [7].

In increasingly interconnected systems and digitised services, safety hazards can emerge from security vulnerabilities. An unsecured software-based system, such as a Supervisory Control and Data Acquisition (SCADA) system in a power plant, might expose operational networks to external malicious attacks, potentially triggering safety-critical industrial accidents or incidents [8]. This reinforces the need for an integrated approach that transcends the accidental and intentional gap, perhaps with 'risk' as the integrating concept. We now accept that safety hazard analysis must proactively consider security threats as potential sources of harm, and cybersecurity methods must analyse the safety implications of any additional security controls.

The boundaries between safety and security are even more blurred for AI systems [9]. With their reliance on large and often interconnected data sources, deep neural networks for example are particularly prone to adversarial attacks. In particular, perception and vision functions, implemented by these models, are vulnerable to manipulation by malicious actors, e.g. fooling the networks by placing small stickers on traffic signs [10], leading autonomous cars to misclassify stop signs and increasing the risk of collision.

Extending AI Safety to Societal Harms

Policy and regulatory documents already emphasise the potential benefits and dangers of the use of AI. This is evident in, for example, the UK's pro-innovation regulatory approach to AI [11]. As outlined in the Policy Paper published in March 2023, this approach recognises the wide range of AI-induced harms, from the potential of the technology to '*damage our physical and mental health*' to its capacity to '*infringe on the privacy of individuals and undermine human rights*'.

Increasingly, the latter two kinds of societal harms, as well as other broader kinds of ethical and political concerns, feature under the umbrella of AI safety [12, 13]. These include algorithmic discrimination in recruitment and in the justice system, racial bias in accessing health and social care services, targeted generation of misinformation by malicious actors, sudden shifts in the job market and overdependence on a handful of powerful AI companies. The concerns go beyond the individual instances and extend to troubling society-wide patterns that could weaken long-established democratic, social and regulatory institutions and norms. These sources of harm are not new. However, the scale and complexity of AI systems exacerbate and entrench these systemic issues. Further, the opaque nature of AI systems, specifically those incorporating deep neural networks, which may be trained and tested using imbalanced and biassed data, are inaccessible and inscrutable by those directly affected by their deployment (e.g. patients or pedestrians) or their representatives (e.g. regulators or advocacy groups). This leads to significant legal, social and ethical responsibility gaps [14].

"The opaque nature of AI systems, specifically those incorporating deep neural networks, which may be trained and tested using imbalanced and biassed data, are inaccessible and inscrutable by those directly affected by their deployment"

"An AI model might be deemed 'safe enough' to use, but its development might raise significant ethical and even legal concerns".

It is also important to consider the concerns in the AI development process itself, i.e. 'upstream harms' [15]. These include unfair working conditions for those collecting and labelling machine learning datasets, the environmental footprint of training and running foundational AI models, and privacy violation around the use of publicly available data. That is, an AI model might be deemed 'safe enough' to use, but its development might raise significant ethical and even legal concerns.

The rapidly growing landscape of AI harms raises a critical question: what is the appropriate scope for AI safety? To illustrate the potential for harm across various contexts, let's consider these realistic scenarios:

A. A generative AI system that spreads fake news, eroding confidence in independent regulators
B. An AI-based sentencing system that produces racially-biassed recommendations
C. An autonomous car that misclassifies stop signs and leads to collision
D. An AI-based clinical-decision support system that weakens trust in human clinical judgement
E. An AI-based health screening system that is at least as performant as current non-AI services but underperforms for patients with dark skin
F. Generative AI content promoted on social media that undermines confidence in election results

These scenarios highlight how AI can cause harm beyond the traditional focus on physical or psychological impacts. There are several approaches to framing the scope of AI safety harms:

1. **Broad Scope**: This approach addresses ethical, legal, social, political, and economic harms (covers all scenarios)
2. **Traditional Focus**: This approach prioritises physical and psychological harm to humans, along with physical damage to property and the environment (only scenario (C) applies)
3. **Refined Focus**: This approach expands option 2 by including additional harms from option 1 that refine existing categories in as far as the nature of the final harmful outcome is physical, psychological or environmental in nature (scenario (E) joins (C))
4. **Direct Causation**: This approach expands on option 3 by including additional harms from option 1 as direct causal factors (scenario (D) joins (C) and (E))
5. **Comprehensive Approach:** This approach builds on option 4 by considering additional harms from option 1 as contextual or indirect causal factors (scenario (A) joins (C), (D), and (E). Scenarios (B) and (F) might also be relevant, although their direct impact is harder to establish)

Option 1, while offering a holistic view, might overwhelm current safety methodologies. Option 2 is too narrow, neglecting crucial aspects. Options 3-5 offer a spectrum of possibilities, with Option 5 seeming the most promising. It maintains a focus on core harms (physical, psychological, environmental) while explicitly addressing equity in safety. Additionally, it considers other potential harms (such as misinformation) as contributing factors, striking a balance between scope and feasibility.

Where do we go from here?

The answer does not lie in artificial boundaries or fragmented efforts. Fundamental safety norms endure: starting early, adopting a whole-system mindset, and maintaining safety through-life remain unchallenged by AI. AI or not, physical and psychological harm to humans and physical damage to property and the environment will rightly remain the focus of safety activities. However, an open and inclusive debate is unfolding on how to integrate other AI-induced harms, especially ethical and legal concerns about bias and discrimination. This should involve methods for assessing algorithmic fairness and meaningfully incorporating diverse stakeholders throughout the AI development process. Whether this is reflected in an expanded scope for safety engineering, or in a new all-encompassing "AI Safety" interdisciplinary profession remains to be seen. Whichever emerges, it is clear that the next generation of professionals will need broader, multi- disciplinary knowledge – something we are seeking to foster in our UKRI Centre for Doctoral Training in AI Safety (SAINTS) [16].

The aim of SAINTS is to train a cohort of 60 PhD students with the research expertise and skills necessary to ensure that the benefits of AI-enabled systems are realised without introducing harm as the systems and their environments evolve. Research will be focused on the lifelong safety assurance of increasingly autonomous AI systems in dynamic and uncertain contexts, building on methodologies and concepts in disciplines spanning AI, safety science, philosophy, law and the social sciences. SAINTS will bring together students and partners from a diversity of backgrounds and sectors to deliver a new generation of experts who make leading contributions to AI safety.

References

[1] Zoe Porter, Ibrahim Habli, John McDermid, and Marten Kaas. "A principles-based ethics assurance argument pattern for AI and autonomous systems." AI and Ethics (2023): 1-24.

[2] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave. "Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research." London: Nuffield Foundation (2019).

[3] Department for Science, Innovation and Technology. "Frontier AI: capabilities and risks – discussion paper." https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper (October 2023).

[4] Sidney Dekker. Second victim: error, guilt, trauma, and resilience. CRC press, 2013.

[5] John Alexander McDermid, Simon Burton and Zoe Porter. "Safe, ethical and sustainable: framing the argument." In The Future of Safe Systems: Proceedings of the 31st Safety-Critical Systems Symposium, 2023, pp. 297-316. Safety Critical Systems Club, 2023.

[6] National Transportation Safety Board. "Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona March 18, 2018", https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf. (2019)

[7] Robin Bloomfield, Kateryna Netkachova and Robert Stroud. "Security-informed safety: if it's not

secure, it's not safe." In Software Engineering for Resilient Systems: 5th International Workshop, SERENE 2013, Kiev, Ukraine, October 3-4, 2013. Proceedings 5, pp. 17-32. Springer Berlin Heidelberg, 2013.

[8] David Kushner. "The real story of stuxnet." ieee Spectrum 50, no. 3 (2013): 48-53.

[9] Department for Science, Innovation and Technology. "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023". https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[10] Nir Morgulis, Alexander Kreines, Shachar Mendelowitz and Yuval Weisglass. "Fooling a real car with adversarial traffic signs." arXiv preprint arXiv:1907.00374 (2019).

[11] Department for Science, Innovation and Technology. "Policy paper A pro-innovation approach to AI regulation" https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper (3 August 2023)

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

[13] Ada Lovelace Institute. "Regulating AI in the UK", https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk (July 2023)

[14] Simon Burton, Ibrahim Habli, Tom Lawton, John McDermid, Phillip Morgan and Zoe Porter. "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective." Artificial Intelligence 279 (2020): 103201.

[15] Ada Lovelace Institute. "Regulating AI in the UK", https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk (July 2023)

[16] UKRI AI Centre for Doctoral Training in Lifelong Safety Assurance of AI-enabled Autonomous Systems (SAINTS): https://www.york.ac.uk/study/postgraduate-research/centres-doctoral-training/safe-ai-training

Imagine Attribution:
leading image: AI generate by Midjourney
Tempe accident: Public Domain: https://commons.wikimedia.org/wiki/File:Tempe_prelim_figure_2_(41583363594).png

Professor Ibrahim Habli, Research Director, Centre for Assuring Autonomy, University of York

A leading expert in design and assurance of safety-critical systems, Ibrahim works closely with industry and regulators and has made significant contributions in dynamic safety cases and safe and ethical use of AI. Driven by a passion for interdisciplinary safety research and team science, he has led major, complex research activities, with responsibility for collaboration with organisations like the NHS and Jaguar Land Rover. He also directs the UKRI Centre for Doctoral Training in AI Safety (SAINTS).

Professor John McDermid, Director, Centre for Assuring Autonomy, University of York

An internationally recognised speaker and expert in system, safety and software engineering, his role at the centre focuses on safety of robotics and autonomous systems. He has acted as an advisor to government and industry for several decades, working in several organisations as a Non-Executive Director, including the Health and Safety Executive (HSE), and as an Advisory Board Member of organisations such as the Ministry of Defence, Rolls-Royce and the Fraunhofer Institute in Kaiserslautern. In 2002 he became a Fellow of the Royal Academy of Engineering and was awarded an OBE in 2010.