Concept Creep in Safe Artificial Intelligence

Laura Fearnley, Ibrahim Habli
University of York
laura.fearnley@york.ac.uk, ibrahim.habli@york.ac.uk

Abstract

This paper argues that the concept "safety" in AI has undergone concept creep, a phenomenon which describes the gradual semantic expansion of harm-related concepts. Originally observed in psychology, concept creep involves concepts broadening their meaning both vertically, to include less severe phenomena, and horizontally, to encompass qualitatively new phenomena. We argue that safety, particularly when applied to AI, has crept along both axes. Our analysis traces this creep by contrasting a baseline definition of safety, which is grounded in the discipline of safety science, with contemporary discourse on the safety of AI systems. We demonstrate that safety has crept horizontally to cover new phenomena, such as systemic injustices and existential risks, and it has crept vertically to include less severe phenomena, such as those related to mental wellbeing. The primary aim of this paper is to map the conceptual expansion of safety. We stop short of arguing whether this expansion constitutes progress or regress for the design and development of AI systems. However, we argue that the process of concept creep produces both beneficial and costly effects for society, policy, industry, and academic research communities. We suggest that some of the promising developments and the problematic trends recently witnessed within AI safety discourse can be understood, at least in part, as a consequence of concept creep.

Introduction

As the advancements in AI proceed expeditiously, questions surrounding their safe development and deployment have become increasingly urgent. AI safety, as a research field, incorporates, amongst other things, technical research on topics like robustness and reliability (Akram et al. 2022), adversarial attacks (Ali et al. 2020), out of distribution detection (Kamoi and Kobayashi 2020), XAI methods (Jia et al. 2022), and analysis of failure modes (Martinez et al. 2023), as well as more theoretical topics including how to develop safety cases for AI systems (Clymer et al. 2024). Despite its centrality to public discussions of AI, no clear consensus has emerged, either in practice or in academia, about what constitutes a safe AI system. Plenty of research papers have attempted to fill this lacuna by defending a definition of safety and thereby delineating what research projects should fall under its purview (Gyevnar and Kasirzadeh 2025; Harding and Kirk-Giannini 2025; Chiodo et al. 2025).

In this paper, we take a different approach to analysing the safety of AI systems. We do not aim to gain normative ground by defending our preferred definition of safety, rather, we take a step back to look at the evolution of the concept and its associated meaning within AI discourse. We find that following technical advancements in AI, what it means for a system to be safe has dramatically shifted. Specifically, we argue that the concept of safety has undergone "concept creep", a phenomenon which describes the gradual semantic expansion of harm-related concepts over time (Haslam 2016). Originally observed in psychology, concept creep involves concepts broadening their meaning both vertically, to include less severe phenomena, and horizontally, to encompass qualitatively new phenomena (Haslam 2016). We argue that safety in AI has crept along both axes.

Our analysis traces this creep by contrasting a baseline definition of safety, which is grounded in the discipline of system safety or safety science, with contemporary discourse on the safety of AI systems. We suggest that safety has expanded horizontally, beyond its historical engineering roots, to cover qualitatively new kinds of phenomena, specifically, systemic injustices and speculative long-term existential risks. Vertically, the concept has broadened downwards; safety in AI is increasingly associated with preventing less severe degrees of harm, such as those related to mental wellbeing.

The creep of safety is not random semantic drift, but driven, amongst other things, by the evolution of AI technology itself. Changes in complexity, adaptability, interactivity, and scale have enabled AI systems to generate and intensify harmful phenomena beyond the capabilities of traditional, non-AI systems. Those employing large language

models (LLMs) or recommender algorithms, can dynamically personalise outputs in ways that reinforce existing cognitive biases or exacerbate emotional vulnerabilities (Chandra et al. 2025), and their capacity to rapidly scale across platforms enables them to propagate these risks at a pace and level of granularity that are beyond the reach of conventional systems. The proposal advanced here is that, as a consequence of these advancements, the concept of safety is creeping to account for the novel and amplified risks posed by AI systems.

One important comment on what's to follow. In this paper, our primary aim is to map the conceptual expansion of safety. We stop short of defending a position with regards to whether this expansion constitutes overall progress or regress for design and development of responsible AI systems. However, we argue that the *process* of concept creep produces both beneficial and costly effects for society, policy, industry, and academic research communities. We suggest that some of the promising developments and problematic trends recently witnessed within AI safety discourse can be understood, at least in part, as a consequence of creep. So, whilst we do not argue that concept creep constitutes progress or regress, we do draw out some significant normative implications of the process.

To develop this argument, we begin by outlining concept creep. Next, we introduce a baseline definition of safety which is rooted in safety science. Then we show that this conception of safety has crept both horizontally and vertically following advancements in AI systems. We end by examining some potential costs and benefits of safety's creep, before reflecting on the trajectory of the field as its conceptual boundaries continue to evolve.

Conceptual Framework: Concept Creep

The backdrop for analysing the shifting conceptions of safety is "concept creep". The term was introduced by psychologist Nick Haslam (2016) to describe the gradual semantic expansion or inflation of harm-related concepts over time, such that they come to encompass a broader range of phenomena than they originally did. Haslam initially evidenced his theory with case studies focused on abuse, bullying, trauma, mental disorder, addiction, and prejudice, arguing that their boundaries had stretched, and meanings dilated over recent decades. However, the theory is not limited to these examples, for it posits a general tendency for harm-related concepts to broaden semantically. Later work explicitly clarified that the phenomenon can apply to positive or desirable concepts associated with mitigating harm, such as care, compassion, and safety (Vylomova et al. 2019).

Haslam (2016) distinguishes two primary mechanisms through which this semantic expansion occurs: vertical creep and horizontal creep. Vertical creep involves the downward extension of a concept's meaning to encompass milder, less extreme, or less severe variants of the phenomenon it originally denoted. This can occur through a lowering of the threshold for identifying the phenomenon or a relaxation of the defining criteria. For example, the concept of bullying was initially used in the 1970s to refer to aggressive behaviour among children that was intentional, repeated, and perpetrated downward in a power hierarchy. Bullying underwent vertical creep to include less extreme behaviour such as acts that were unintentional, unrepeated, and directed at people of equal or higher power than the perpetrator (Haslam et al. 2021). Horizontal creep, in contrast, involves the outward extension of a concept to encompass qualitatively new phenomena or its application in entirely new contexts. For instance, bullying moved horizontally to include the behaviour of adults in workplaces, exclusionary rather than intimidating behaviour (e.g., shunning), and intimidation carried out online rather than in-person ("cyberbullying"). Concept creep is largely considered to be ambivalent (Haslam 2016; Furedi 2016). On the one hand, broadening harm concepts allows for the recognition of previously overlooked forms of suffering and maltreatment (Cascardi and Brown 2016). On the other hand, concept creep risks pathologising everyday experiences at the cost of trivialising more severe instances through semantic dilution (Harper et al. 2023).

Since Haslam's 2016 paper, the theory has gained wide recognition in psychology and beyond, prompting a substantial body of empirical research (Brandt and Proulx, 2016; Niemi and Young 2016; Haidt 2016). Various drivers of concept creep have been proposed. Some suggest it reflects a growing cultural sensitivity to risk and vulnerability in Western societies (Furedi 2016), which has led to the identification and problematisation of forms of suffering and maltreatment that were previously overlooked or tolerated. Relatedly, the decline in violence in the West (Pinker 2011) may have led to the inclusion of milder or borderline instances of harm within existing concepts that were reserved for more severe phenomena. Other factors also play a role in creep, including the emergence of new phenomena requiring conceptualisation (e.g., cyberbullying), and political actors strategically broadening concepts to amplify the perceived seriousness of a social issue (Charmaz et al. 2019).

Conceptualisations of safety have been shaped by some of the dynamics and drivers that parallel those which have led to concept creep in psychology. AI technology has introduced novel harmful phenomena, such as "toxic content" and deepfakes (Weidinger et al. 2022), all of which require conceptualisation. Furthermore, the potential for AI systems to affect our lives and livelihoods on unprecedented scales may have led to a heightened sensitivity to risk among researchers, the government, and the public. This confluence of new risks and moral concern creates fertile ground for the

semantic expansion of safety beyond its traditional boundaries, mirroring the psychological patterns Haslam described. The expanded meaning of safety which we wish to highlight here, may therefore not be mere semantic drift, but a process actively shaped by the evolving landscape of artificial intelligence. [1]

Conceptions of Safe AI: Foundations in Safety Science

To see that safety has crept, we require a baseline against which subsequent conceptual expansion can be measured. To this end, we take as our starting point a conception of safety articulated in the field of safety science. The safety science conception serves as a germane baseline for three chief reasons.

First, many have argued that safety research in AI is an extension of this broader tradition. Different authors make different claims here; some contend that the historical roots of contemporary AI safety can be traced to systems engineering and safety science practices (Gyevnár and Kasirzadeh 2025), while others characterise AI safety as a special case of safety engineering (Hendrycks 2024). Several have also advanced the view that ongoing research in this area should align itself with the discipline of safety science. For instance, Dobbe (2022) advocates for continuously applying insights from safety science to AI development and deployment, while Weidinger et al. (2023) defend the claim that AI safety research should be fully subsumed within the discipline of safety science. Regardless of the particulars, there is a clear trend framing AI safety as closely aligned with safety science. In practice, this alignment is reflected in the objectives, methodologies, and conceptual frameworks that the AI safety research community has inherited. Much technical and non-technical work in this area relies on established safety science methods to evaluate and manage the risks posed by AI systems (Rismani et al. 2023; Hendrycks et al. 2021; Johnson 2022). This historical and ongoing alignment makes safety science a natural foundation for our initial conceptualisation of safety.

Second, the safety science conception of safety is operationalised in legally and industrially recognised standards. For example, the international standard IEC 61508, which governs the functional safety of electrical, electronic, and programmable electronic safety-related systems, uses a conception of safety that is both informed by and reflected in safety science discourse. The definition is similarly embedded across its sector-specific derivatives, such as ISO 26262 for the automotive industry. These international standards provide frameworks for designing, implementing, operating, and maintaining safety-related systems, many of which govern the kinds of systems into which AI is being increasingly embedded. Compliance is widely regarded as best practice and often serves as evidence that legal and regulatory requirements have been met (IEC and ISO 2015). Using the safety science definition as a baseline thus grounds our analysis in the realities of engineering practice and regulatory compliance.

The third reason to begin with safety science is pragmatic. It turns on the thought that safety has a literal meaning within the field. There is general consensus on what safety means, and what conditions must be met to develop safe systems, albeit at a high-level of description. Safety science therefore provides a lucid notion of safety which can serve as a baseline against which subsequent conceptual expansion can be examined. For the purposes of our analysis then, we take our initial understanding of safety as one inherited from safety science.

What, then, is the safety science understanding of safety? Safety is often characterised as the state of being free from, or not causing, unacceptable risk of harm (Lowrance 1976; Habli 2025a; Knight 2002; Bozzano and Villafiorita 2010; IEC 2010). It's important to stress that this is by no means the only way of understanding safety in the field (Swuste et al. 2011; Hollnagel et al. 2015), but it is a dominant definition, especially in industry. It's also important to recognise that safety is a context sensitive property. Whether a system is safe depends upon the operational context and the system's ability to carry out its intended purpose within and beyond the defined context. Two key principles underpin our adopted conceptualisation of safety that are worth delineating.

Risk acceptability: Risk is typically defined as a combination of the likelihood and severity of harm (Leveson 2023). Complete elimination of risk is often impossible or impractical. Whether a system can be classified as safe therefore depends upon whether the level of risk posed by the system is acceptable. Consequently, risk assessments form a foundational part of safety methodologies, and include the identification, analysis and evaluation of risk. There is no universal standard of what constitutes an acceptable level of risk. Ultimately, decisions about acceptability depend on several criteria which change according to the industrial sector and the specific application at hand. Decisions about acceptability entail value judgements based on existing good practices (Manuele 2010).

[1] Haslam's (2016) account focuses on the inflation of harm-related concepts (e.g., trauma, bullying, prejudice) over time. Our use of concept creep similarly reflects on the semantic expansion of safety in light of technological advancements over time, but it also considers changes following contextual shifts between safety in general and AI safety in particular.

Harms: Safe systems do not cause unacceptable risk of harm. According to the safety science definition adopted here, not all kinds of harm produced by a system constitute compromises to the safety of that system. The kinds of harm that can compromise safety are those that affect **human health, property**, and **the environment** (Knight 2002; Bozzano and Villafiorita 2010; Manuele 2010; Bell 2006).

The types of harm traditionally prioritised in safety assessments are those with direct, immediate, and often physical manifestations. In terms of human health, emphasis has typically been placed on bodily damage and severe psychological injury, with death, breaks, burns, and trauma serving as core categories of concern. For instance, a system that poses an unacceptable risk of causing traumatic brain injury would be categorised as unsafe under conventional safety standards. Moreover, safety researchers also address a system's ability to reduce cognitive function through cognitive overload, fatigue, and distraction caused by overly complex or poorly designed interfaces (Wickens et al. 2021). In aviation, for example, a pilot's performance may deteriorate under high cognitive load induced by information dense cockpit displays, potentially leading to lapses in attention and delayed response times. In addition to concerns about human health, safety assessments have traditionally focused on damage to property, including the loss or degradation of infrastructure, critical resources, and equipment. Environmental harm, such as ecological degradation, pollution, and habitat destruction, have also become increasingly recognised as a relevant dimension of safety, particularly within high-impact sectors like transport and energy (Habli 2025a).

In the next section, we argue that the concept of safety has expanded both horizontally to encompass new categories of harm, and vertically by lowering the threshold for what qualifies as a relevant harm. Our analysis focuses on how concept creep has reshaped the core harm domains traditionally associated with safety: human health, property, and the environment.

The safety science conceptualisation of safety outlined here has evolved organically through discussions among a diverse range of stakeholders, including customers, regulatory bodies, user groups, insurance companies, professional societies, and industry standards organisations (Leveson 2023). The focus on harm to human health, property, and the environment arguably reflects the nature of the systems that safety science has traditionally addressed. Non-AI systems are limited in both complexity and societal reach, which in turn constrains the kinds and degrees of harm they could produce. As the systems under evaluation grow more complex and socially embedded, conceptualisations of safety shift too. In the next section, we examine how the changing landscape of systems has changed the notion of safety itself as it moves to absorb a wider variety of phenomena that were once thought to fall outside its scope.

Safety's Creep

The initial conception of safety as grounded in the field of safety science formed at a time when systems were simpler and less interdependent. Hollnagel et al. (2015) explain that in traditional systems, the dependence on Information Technology (IT) was limited (mainly due to the size and immaturity of IT itself), which meant these systems largely operated according to fixed rules and predictable processes, making it possible to understand and follow, to a high degree of confidence, what went on in a system. Furthermore, the level of integration across systems and sectors was low (Hollnagel et al. 2015), systems were typically static and domain specific. Importantly, this meant that traditional systems were naturally limited in what kinds of phenomena they could produce and the degree to which they could impact individuals and society. For example, in the context of nuclear power, safety protocols were developed around a well-defined set of physical and procedural processes, with clear thresholds for failure and well-understood causal mechanisms. Nuclear power systems were largely closed and their boundaries well-specified, primarily making safety assessment a matter of rule-following and technical redundancy rather than adaptive or cross-domain risk management (National Research Council 1997).

However, significant changes have occurred in the types of systems being built today and the context in which they are being deployed. Contemporary AI systems are dynamic, often unpredictable, largely opaque, and have a high level of integration across subsystems and sectors. Those incorporating frequently updated machine learning components, can adapt and learn over time (e.g. reinforcement learning from human feedback), producing emergent and unpredictable behaviours. Foundational models (e.g., OpenAI's GPT-4, Anthropic's Claude) present new levels of complexity and capabilities, performing a wide range of general tasks, including text synthesis, image manipulation and audio generation. Their ability to engage in open-ended conversations across a wide variety of domains with fluency and contextual sensitivity marks a qualitative shift in how systems can communicate, respond, and influence users. Large-scale deployment across sectors and higher levels of integration have also given AI systems further reach than their traditional counterparts. Changes in complexity, adaptability, interactivity, and scale have meant that AI systems can produce novel kinds of harm that were improbable with older, non-AI technologies, while also amplifying familiar harms to degrees not previously addressed in traditional safety science.

The proposal advanced here is that as a direct consequence of technological advancement, the concept of safety has undergone a process of conceptual expansion as it absorbs a wider variety of phenomena that were once thought

to fall outside its scope. In the next section, we turn to examine one particular mode of this expansion—horizontal creep—which involves the inclusion of qualitatively different *kinds* of harm under the umbrella of safety. This form of creep has significant implications for how safety in AI is operationalised and governed.

Horizontal Creep

In this section, we argue that what is meant by "safe" AI systems has undergone horizontal creep. This involves the broadening of its scope to encompass qualitatively new kinds of phenomena that extend beyond the traditional safety science focus on human health, property, and the environment. Arguably, there have been several new phenomena folded into the concept of safety in AI. In this section, we focus on two particularly significant additions: systemic injustices and existential risks.

Perhaps the most salient expansion is the increasing inclusion of systemic injustices within safety discourse. Systemic injustices or systemic harms refer to broad, widespread negative impacts that extend beyond individuals to affect entire communities, societal structures, or ecosystems. Philosopher Sally Haslanger (2023) argues that systemic injustice occurs when "an unjust structure is maintained in a complex system that its self-reinforcing, adaptive, and creates subjects whose identity is shaped to conform to it". Unjust societal structures can encode patterns like unjustified biases, racial and gender discrimination, and their manifestations can undermine the attainment of long-established goods and norms, such as democratic institutions, principles of justice, human rights, and personal autonomy. The ways in which AI systems can contribute to the maintenance of unjust structures, thereby producing systemic harms, is becoming a prominent focus within the conversation about the safety of such systems. For instance, the UK Government's AI Safety Institute (now rebranded as the AI Security Institute) argues that "safety-relevant properties" include "future societal harms" which can manifest through a system's "psychological impacts, its capacity for manipulation and persuasion, its influence on democracy, biased outputs and reasoning, and systemic discrimination" (UK Government, Department for Science, Innovation and Technology, 2024). Similarly, the 2023 AI Safety Summit, a global event that brought together governments, industry, academia, and civil society, published a discussion paper, where they argued that safe AI systems should be assessed according to three broad categories: societal harms, misuse, and loss of control. Under societal harms, the authors identified issues such as bias, fairness, representational harms, and disruptions to labour markets (UK Government, Department for Science, Innovation and Technology, 2023).

Beyond government initiatives, independent research organisations have also called for a broader understanding of safety in AI. The Ada Lovelace Institute, in its 2023 report 'Regulating AI in the UK', argued that "[i]t will be important for the definition of 'AI safety' used by the Government, the Foundation Model Taskforce and the AI Summit to be an expansive one, reflecting the wide variety of harms that are arising as AI systems become more capable and embedded in society" (Davies and Birtwistle 2023). The report categorises AI safety-related harms into four types: accidental harms from system failures or unexpected behaviours (e.g., self-driving car crashes or discriminatory hiring algorithms); misuse by bad actors (e.g., the spread of misinformation through generative AI); structural harms from changes to social, political, or economic dynamics (e.g., the erosion of democratic institutions due to widespread misinformation); and upstream harms arising further up the AI value chain (e.g., poor labour practices).

In an article published in *Science*, Alondra Nelson, who spearheaded the White House Blueprint for an AI Bill of Rights, alongside co-author and philosopher Seth Lazar, neatly encapsulated the shift toward including systemic injustices in AI safety management. They wrote: "Years of sociotechnical research show that advanced digital technologies, left unchecked, are used to pursue power and profit at the expense of human rights, social justice, and democracy. Making advanced AI safe means understanding and mitigating risks to those values, too" (Lazar and Nelson 2023). The discourse reveals that the locus of safety-relevant phenomena has expanded from our initial starting point within safety science to include new phenomena relating to the design of equitable and just systems.

One important clarification is needed here to ward off a potential objection. We do not wish to insinuate that our initial baseline for safety as one aligned with safety science was not at all concerned with preventing systemic injustices. Rather, the point we wish to highlight is that preventing the unacceptable risk of certain immediate harms (i.e. to human health, property and the environment) were, and by many still are, considered to be the core categories of concern for the safety community. Other kinds of phenomena, such as systemic injustices, were treated, if at all, as peripheral, contributory or secondary safety concerns.

The second major horizontal expansion we wish to highlight here involves incorporating concerns about long-term, catastrophic, existential risks (often termed "x-risks"). X-risks are generated by concerns regarding the potential capabilities of future, highly advanced AI systems, such as Artificial General Intelligence (AGI) or Artificial Superintelligence (ASI). This perspective posits that sufficiently intelligent AI systems might develop goals misaligned with human values, become power-seeking, resist shutdown, or otherwise pose uncontrollable threats leading to human extinction or irreversible collapse of civilisation. Research agendas within this area focus on technical problems like value

alignment, scalable oversight, and detecting deceptive behaviour, as well as normative work connected to movements such as rationalism, effective altruism, or longtermism (Gyevnar and Kasirzadeh 2025).

X-risks are becoming increasingly interwoven into what constitutes a safety concern for AI systems. Many AI researchers, major AI labs, and organisations like the Centre for AI Safety are explicitly linking AI safety with the mitigation of such existential threats. A recent study of the 'epistemic community' of AI safety characterises the discipline as follows: "generally, AI safety practitioners are interested in preventing catastrophic long-term events precipitated by the deployment of machine learning systems" (Ahmed et al. 2024). The European Network for AI Safety (ENAIS), a consortium of experts, researchers, and policymakers, explains that "AI safety is about the safeguarding of humanity from uncontrollable AI scenarios. This can include global systemic risks like nuclear war and cyberwarfare from the use of artificial intelligence but also the dangers from self-improving AI systems with their emergent uncontrollable goals" (ENAIS 2023). Elsewhere Richard Ngo, an AI researcher who has worked at OpenAI and DeepMind, writes that the key concern motivating safety research on AGI is that "if they [AIs] don't want to obey us, then humanity might become only Earth's second most powerful "species", and lose the ability to create a valuable and worthwhile future" (Ngo 2020).

Gyevnar and Kasirzadeh (2025) observe that contemporary discourse on the safety of AI systems has increasingly been framed as a project aimed at minimising existential risks associated with future, highly advanced AI systems. They lament that "[t]his concentrated attention on existential risk has emerged despite — and perhaps overshadowed — decades of engineering and technical progress in building robust and reliable AI systems".

As the preceding discussion indicates, x-risks encompass more than the threat of human extinction. They also include concerns about the erosion of human agency, autonomy, and self-authorship. In this sense, x-risks go beyond traditional safety concerns about physical harm, extending to worries about humanity's capacity to create value and meaning. The focus on speculative, high-impact, long-term risks therefore constitutes a horizontal expansion to the safety agenda, introducing a qualitatively distinct category of concern.

Vertical Creep

In the previous section, we argued that safety in AI has come to encompass qualitatively new kinds of phenomena, stretching it beyond the traditional safety science focus. In this section, we argue that safety in AI has also come to encompass *degrees* of phenomena that extend beyond its initial scope. The expansion in degrees mirrors what Haslam (2016) describes as vertical creep; the downward expansion of a concept to include less severe or extreme phenomena.

The most significant downward expansion, which will be our focus, is the lowering of the threshold for what constitutes harm to human health. According to our initial understanding of safety, safe systems are those that do not cause unacceptable risk of harm to human health, property, or the environment (Knight 2002; Bozzano and Villafiorita 2010; Bell 2006). Within this definition, human health is typically understood in terms of physical integrity and high-level psychological integrity. Injury, death, and severe psychological damage, such as trauma and cognitive burnout, represent the degrees of harm that could compromise the safety of a system (Habli 2025a). Phenomena falling below this threshold, though potentially serious, were once thought to fall outside safety's purview. One barrier to adopting a more expansive notion of human health is that psychological harm is not defined by current laws and regulations, making it profoundly difficult to determine which risks cause psychological harm and, by extension, what forms of regulation could reduce exposure to such risks (Osman 2025). However, as AI systems grow more sophisticated and deeply embedded in our social world, it has become clear that they can impact our health in more subtle and pernicious ways. As a result, there is increasing attention being paid to the ways in which AI systems can impact mental and emotional wellbeing more generally (i.e., influences on affective states like attitudes, moods, emotions, as well as cognitive states).[2]

AI systems, particularly GenAI and LLM driven conversational chatbots, possess unique capabilities for social mimicry, allowing them to role-play with users and produce dialogues that are seemingly deep and meaningful. Empirical research on such systems have shown that in some cases regular use can lead to AI-induced emotional distress, overdependence, and addictive behaviours (Gabriel et al. 2024; Freitas 2024). One study found a deterioration in mental state in over 34.4% of participants that regularly engaged in dialogues with character-based chatbots, and that psychological deterioration was especially prevalent in vulnerable users (Qiu et al. 2025). Other studies have shown that overreliance on chatbots is both a risk factor and a complication of depression (Lock 2023) and loneliness (Laestadius et al. 2022). Beyond direct interactions with GenAI and LLMs, AI-driven systems embedded in social media platforms also present risks to mental health. Recommendation algorithms, engagement-optimised content delivery, and personalised feed generation have all been shown to shape users' emo-

[2] We do not mean to trivialise these effects by characterising them as less severe. We merely wish to convey that thought that some phenomena fall below the threshold for what constitutes a relevant safety phenomenon according to our baseline concept.

tional states through mechanisms such as algorithmically reinforced social comparison, manipulation, and the creation of compulsive interaction loops (Berryman et al. 2018; Montag et al. 2021). These systems exploit cognitive biases and emotional sensitivities to maximise engagement, but in doing so, they can contribute to increased anxiety and feelings of isolation (Twenge et al. 2018).

Following these findings, researchers and regulators working in safety have extended conceptualisations of human health to include the protection of individual emotional and mental wellbeing. For example, Qiu et al. (2025) highlight the potential for "mental health hazards in human-AI interactions," that can result in a failure to uphold essential safety principles. Whilst Zeng et al. (2024) argue that safety benchmarks for AI systems should include categories such as offensive language, hate speech, and sexual content, to mitigate potential impacts on mental wellbeing. Similarly, Zhang et al. (2024) argue that assessing the safety of LLMs necessitates evaluating their impact on users' "emotional wellbeing" and propose the inclusion of a "mental health" benchmark as part of AI safety assessments. Other authors similarly draw direct connections between safety and the risks to mental and emotional wellbeing (Balesni et al. 2024; Osman 2025).

Evidence of vertical creep can also be found in AI safety regulatory frameworks. Article 5 of the EU AI Act, for instance, ties together concerns about psychological wellbeing and AI's potential for manipulation. The Act prohibits AI systems deploying "subliminal techniques beyond a person's consciousness" or that exploit "any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability" in a manner "that causes or is likely to cause that person or another person physical or psychological harm." (EU AI Act. Article 5(1)). The Act is not explicit in its definition of psychological harm. But scholars have argued that it should be understood broadly to include impacts on general mental health to accommodate for the complexity of applying the term 'psychological harm' to the various uses of AI (Pałka 2024).

The particular concern for vulnerable users in discussions of emotional and mental wellbeing suggests a nuanced driver for this vertical creep. It is not simply a generalised rising sensitivity to harm (Furedi 2016), but rather a targeted ethical concern for specific populations perceived as less resilient to the subtle psychological impacts of AI. Risk management techniques in traditional safety science typically consider the risks systems impose on the general public, although specific vulnerable groups (e.g., children and toy safety standards) are also recognised. However, in the AI domain, the capacity for personalised interaction and the potential for subtle forms of manipulation or emotional influence might render certain psychological vulnerabilities more exploitable or exacerbate existing conditions. The recognition of these specific vulnerabilities in human-AI interaction propels the safety boundary downwards to include phenomena which might be considered mild for the general population but can be significant for vulnerable groups. This targeted sensitivity contributes directly to the vertical expansion of safety within AI.

Table 1 presents examples of both the vertical and horizontal creep of safety in AI.

Type of Creep	Phenomena	Description/Example
Horizontal	Systemic injustice: erosion of democratic processes	AI generating and disseminating convincing fake news or propaganda, influencing public opinion, sowing discord, or undermining trust in democratic institutions (UK Government, Department for Science, Innovation and Technology, 2024; Lazar and Nelson 2023).
Horizontal	Systemic injustices: algorithmic bias	AI systems used in hiring, loan applications, or criminal justice making decisions that unfairly disadvantage individuals based on protected characteristics (race, gender, etc.) due to biased training data or model design (Weidinger et al. 2023; Davies and Birtwistle 2023).
Horizontal	Existential risk: loss of control	Superintelligent AI pursues goals misaligned with human values and exerts control over global infrastructure, decision-making, or weaponry in ways that lead to irreversible catastrophe (Ahmed et al. 2024).
Vertical	Psychological deterioration	Conversational AI engaging in inappropriate or manipulative dialogue, leading to user anxiety, depression, or exacerbation of mental health conditions (Qiu et al. 2025).
Vertical	Diminished emotional wellbeing	AI-driven social media algorithms promoting addictive use patterns, social comparison, or exposure to distressing content (Berryman et al. 2018; Montag et al., 2021).

Table 1. Examples of creep in Safe AI.

Consequences of Creep

Thus far, we have aimed to map some of the corners of safety's conceptual expansion. However, we have stopped short in arguing that the expansion constitutes progress or regress for the responsible design and development of AI systems. In line with Haslam's (2016) characterisation, we understand concept creep as an ambivalent outcome, neither positively nor negatively valanced. With that said, we do argue that the *process* of concept creep gives rise to consequences that can be both beneficial and costly to society, industry, policy, and academic research communities. In this section, we suggest that some of the promising developments and the problematic trends recently witnessed within AI safety discourse can be understood, at least in part, as a consequence of concept creep. We begin by outlining two positive developments.

Holistic Risk Management

Risk management methodologies are central to determining whether the risk posed by a system is acceptable and therefore whether the system is safe. Risk management techniques, which include the identification, analysis and evaluation of risk, are supported by standards, guidelines, and best practices which have gradually evolved over decades by the relevant academic disciplines and professional communities (Habli 2025a). Collectively, these techniques, and ones like them, are used with the aim of eliminating or reducing the risk of harm caused by sociotechnical systems. Completing these steps successfully and transparently is essential for well-informed, responsible decision-making on system development and deployment (Stilgoe et al. 2017).

Whilst risk management methodologies have been relatively effective for assuring the safety of non-AI systems, many authors have urged for risk management techniques that take a more holistic approach in the domain of AI. For example, Weidinger et al. (2023) identify a critical "sociotechnical gap" in current AI risk assessments, where AI safety is often evaluated narrowly, focusing predominantly on technical components such as data quality, model architecture, and sampling strategies. While these technical aspects remain vital, they are insufficient on their own to determine the overall safety of an AI system. Instead, an effective approach must integrate human and systemic factors that co-determine risks of harm (Weidinger et al. 2023).

One significant potential advantage of broadening the concept of safety is the enablement of a more holistic and comprehensive approach to risk management. Expanding safety, both horizontally and vertically, encourages practitioners and policymakers to consider the diverse and multifaceted impacts AI systems can have throughout their entire lifecycle, and there is now a burgeoning literature which interlaces a rich landscape of risks relevant for safety evaluations (Zeng et al. 2024; Li et al. 2024; Shelby et al. 2023; Raji and Dobbe 2023).

The benefits of this shift towards more holistic risk management are manifold. In a straightforward sense, it acknowledges the plurality of ways AI systems can induce harmful phenomena, a recognition which is essential for assuring the safety of such systems. Furthermore, risk management not only sheds light on, predicts, and quantifies the likelihood of potential downstream harms, but also surfaces the intricate factors and mechanisms influencing their occurrence. A more comprehensive risk management strategy will thereby cultivate a better picture of how new and familiar risks can interact and coalesce. This upshot is not only beneficial from the point of view of assuring safety, insights can also assist other disciplines, such as AI ethics, which also seek to mitigate the harmful impacts of AI systems. Relatedly, more holistic risk management frameworks will help bring to light the different kinds of normative trade-offs that arise as AI systems are developed and deployed in real-world settings. By performing these functions, holistic risk management frameworks play an important role in responsible innovation and deployment of AI systems.

Safety Cases

A second benefit of safety's conceptual expansion lies in its impact on the development of safety cases, particularly through improved clarity and communication. As the scope of safety broadens, stakeholders are increasingly required to specify what kind of phenomena they aim to mitigate. This demand for specificity ultimately strengthens safety discourse and decision-making by making underlying assumptions and priorities more transparent.

A safety case is a structured argument, supported by explicit evidence, that explains why a system is acceptably safe for a specific application within a particular context (UK MoD 2017). It represents established best practice across various safety-critical sectors, especially in transport and energy, underpinned by regulatory and industry standards (Sujan et al. 2016). The concept originated in the UK nuclear industry over six decades ago. A key overarching goal of a safety case is to promote transparency and improve communication among the many and diverse stakeholders who are either interested in or directly affected by the risks posed by a system.

Current safety literature is largely dominated by arguments focused on the risk of physical harm (Habli et al. 2025b). This emphasis results from the priorities of major safety standards in sectors such as automotive and defence, which traditionally focused on risks to physical health or property. However, more recent work is beginning to broaden the scope of AI assurance to include ethical issues (Porter et al. 2024; Burr and Leslie 2023) and existential

risks (Clymer et al 2024), particularly those associated with general-purpose models. These efforts build on earlier, albeit limited, attempts to expand the remit of safety cases to address safety and security risks in a more integrated fashion (Bloomfield et al. 2013).

The broadening of safety's conceptualisation in the current AI debate presents an opportunity for safety cases to evolve. They can encourage developers and deployers to articulate more explicitly the specific types of harms they consider, and to present clearer arguments for why these risks are relevant and significant to their AI systems. Further, they can explain how such risks have been mitigated to acceptable levels, and for whom. This approach and mindset has the potential to improve clarity and communication, whether the risks involve refinements of existing harms (i.e. vertical creep, such as addressing previously under-recognised psychological wellbeing) or new harms (i.e. horizontal creep, such as addressing the spread of misinformation).

Given that safety is rarely absolute, the central focus of a safety case lies in proportionality and trade-offs (Lowrance 1976). The explicit consideration of new (horizontal) types of harm or refined (vertical) types of harm reinforces the communicative role of the safety case in explaining and justifying how risk-benefit analyses are conducted. It also clarifies why emphasis may be placed on certain risks over others, how trade-off decisions are made, and by whom.

Internal Disciplinary Conflicts

With that said, safety's creep is not without its challenges. In recent years, the field of AI safety has experienced growing internal tensions, with disciplinary disputes and the emergence of distinct factions among those working on different aspects of safety. A particular point of contention seems to be the horizontal expansion which extends safety to include systemic injustices. As an illustrative example, consider the comments made by John Tasioulas, Director of the Institute for Ethics and AI at the University of Oxford, in the wake of the 2023 AI Summit: "as anticipated, the concept of safety is stretched in the Declaration to include not only avoiding catastrophe or threats to life and limb, but also securing human rights and the UN Sustainable Development Goals etc. Pretty much all values under the sun" (University of Oxford 2023). Tasioulas's thoughts were echoed by Reed Albergotti, technology journalist and founder of Semafor, who argued that "AI safety is becoming an umbrella term that lumps nearly every potential downside of software automation into a single linguistic bucket". Albergotti notes that in more traditional industries, we deal with safety very differently: "The Occupational Safety and Health Administration (OSHA), for instance, is tasked with making workplaces safe from physical harm. Imagine if OSHA were also responsible for preventing workplace discrimination, retaining workers who are laid off [...] that's similar to what some people are suggesting we do with AI Safety" (Albergotti 2024). Tasioulas and Albergotti express two related worries here. Firstly, incorporating wider systemic injustices into safety assessments would in practice require mitigating against a whole host of undesirable outcomes, thereby significantly overloading the discipline. And secondly, that this overload risks turning AI safety into a nebulous, ill-defined concept. These debates reflect broader uncertainties about the aims and limits of AI safety and raise important questions about how best to define and manage the risks posed by increasingly complex sociotechnical systems.

There is also a backlash with regards to the inclusion of x-risks in AI safety research. The focus on x-risks has been particularly connected to normative theories and movements such as effective altruism or longtermism. Although these theories offer valuable perspectives on long-term challenges, their specific institutional or industrial embodiments have attracted considerable criticism (Acemoglu 2024). This development has been concerning for researchers and practitioners who work on safe and responsible AI, but do not wish to align themselves with x-risk-related normative movements. The inclusion of x-risks has led to a somewhat divided community, with some researchers even questioning the contribution of the "AI safety" community (Bender 2023; Albergotti 2024).

Whether members of the AI safety community lament or welcome the conceptual expansion of safety, it has nonetheless contributed to a somewhat fractious atmosphere within the field. Indeed, the field of AI safety is increasingly defined not only by its subject matter but also by recurring internal disagreements over its scope and priorities. This ongoing contestation reflects deeper tensions about what kinds of phenomena count as safety-relevant, who gets to define them, and which methodologies are deemed authoritative.

Discussion

The scope of safety in AI is not merely of semantic interest; it is inherently political. As noted in the original outline of concept creep, the expansion of concepts can be influenced by underlying moral or political agendas (Haidt 2016; Charmaz et al., 2019). In the case of AI, how safety is conceptualised has far-reaching consequences for regulation, policy, research, and the formation of academic communities. Disciplinary boundaries determine which risks are prioritised, how funding is allocated, which areas of expertise are valued, and the kinds of regulation that are pursued. Harding and Kirk-Giannini (2025) explain that disciplinary boundaries influence every aspect of research, from what researchers are expected to read, who they collaborate with, how research is evaluated, and which directions gain intellectual and institutional traction. Boundaries also shape how disci-

plines interact with the wider world, affecting who is included in public and policy debates, who is responsible for assuring safety, and who regulates such assurances. Conceptualisations of safety are thus pivotal. For these reasons, further analyses of safety's creep requires not only mapping semantic shifts but also understanding the underlying political drivers. Safety's creep is both a symptom of a field which is grappling with a complex, fast-moving and unpredictable technology, and a tool in ongoing struggles over resources, influence, and the contested future direction of AI safety.

In addition to affecting the field's own boundaries, we might wonder how concept creep impacts the relationship with other fields. Conceptual expansion inevitably blurs the boundaries between AI safety and related fields like AI ethics and AI governance. Is ensuring algorithmic fairness primarily a matter of safety or ethics? Is preventing AI-driven erosion of democratic norms a safety concern or a matter of political governance? In many cases, the answer may not be exclusive; rather, these challenges straddle multiple domains simultaneously, reflecting the inherently interdisciplinary nature of the risks posed by advanced AI systems. With that said, maintaining some domain distinctions might be valuable. A distinction, even if imperfect, could help allocate forward-looking responsibilities, identify backward-looking responsibilities, tailor methodologies, and structure research agendas. Furthermore, completely collapsing disciplinary boundaries might lead to problematic implications for AI design and research. For instance, it has been argued that conflating safety properties with other valuable properties may obscure what kind of benchmarks are needed to assess safety and hinder the ability to evaluate trade-offs between assuring system safety and achieving other valuable goals (Ren et al. 2025). At its worst conflating properties can enable "safetywashing"—misrepresenting or misinterpreting improvements in the general capabilities of the AI model as safety advancements (Ren et al. 2025). These risks suggest that the conceptual expansion of safety should be critically examined and strategically managed. Examining safety's semantic expansion is not merely an academic exercise, its boundaries have important implications for how AI systems are designed, evaluated and regulated.

Conclusion

This paper has argued that the concept of "safety" in the context of Artificial Intelligence has undergone a process of "concept creep". Drawing upon safety science as a baseline—where safety is often characterised as the prevention of unacceptable risk of harm to human health, property, and the environment—the analysis demonstrates that safety has expanded both vertically and horizontally. Vertical creep is evident in the increasing inclusion of less severe phenomena, such as emotional and mental wellbeing, and horizontal creep is marked by the incorporation of qualitatively new categories, including systemic injustices and existential risks.

We have suggested that as a direct consequence of technological advancement in AI systems, our baseline concept of safety has undergone a process of conceptual expansion as it absorbs a wider variety of concerns and phenomena that were once thought to fall outside its scope. The semantic expansion of safety may be overall ambivalent, but the process can have both profound costs and benefits to research, policy, and the regulation of AI technology. The ultimate challenge posed by the conceptual expansion of safety lies in striking a delicate balance. It is crucial to acknowledge and address the full spectrum of risks that AI systems can create and perpetuate, yet this must be done in a way that results in practical, actionable, and conceptually coherent frameworks. An "everything is a safety issue" approach risks rendering the concept of safety so broad as to become operationally meaningless. Conversely, an overly narrow definition, which clings strictly to traditional safety science paradigms, may well fail to address many of the real and pressing negative consequences emerging from the deployment of advanced AI. The future necessitates a more nuanced and open approach, whereby researchers, practitioners, and policymakers are explicit about their own conceptualisations of safety.

Acknowledgements

This work was supported by the Centre for Assuring Autonomy, a partnership between Lloyd's Register Foundation and the University of York.

References

Acemoglu, D. 2024. The AI Safety Debate is All Wrong. Project Syndicate. https://www.project-syndicate.org/commentary/ai-safety-human-misuse-more-immediate-risk-than-superintelligence-by-daron-acemoglu-2024-08. Accessed: 2025-08-05.

Ahmed, S.; Jaźwińska, K.; Ahlawat, A.; Winecoff, A.;Wang, M. 2024. Field-Building and the Epistemic Culture of AI Safety. *First Monday*, 29(4). doi.org/10.5210/fm.v29i4.13626.

Akram, M.N.; Ambekar, A.; Sorokos, I.; Aslansefat, K.; Schneider, D. 2022. StaDRe and StaDRo: Reliability and Robustness Estimation of ML-Based Forecasting Using Statistical Distance Measures. In Proceedings of the Computer Safety, Reliability, and Security (SAFECOMP). Cham: Springer. doi.org/10.1007/978-3-031-14862-0_21

Albergotti, R. 2024. The Risk of Expanding the Definition of 'AI Safety''. https://www.semafor.com/article/03/08/2024/the-risks-of-expanding-the-definition-of-ai-safety Accessed: 2025-08-01.

Ali, M.; Hu, F, Y.; Luong, D, K.; Oguntala, G.; Li, J.; Abdo, K. 2020. Adversarial attacks on AI based intrusion detection system for heterogeneous wireless communications networks. In Proceedings of AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). doi.10.1109/DASC50938.2020.9256597.

Bloomfield, R.; Netkachov, K.; Stroud, R. 2013. Security-informed safety: if it's not secure, it's not safe. In Proceedings of Software Engineering for Resilient Systems: 5th International Workshop, SERENE. Berlin: Springer. doi.org/10.1007/978-3-642-40894-6_2

Bell, R. 2006. Introduction to IEC 61508. In Proceedings of the 10th Australian workshop on Safety critical systems and software (SCS '05), Vol. 55. Australia: Australian Computer Society, Inc.

Bender, E. M. 2023 Talking about a 'schism' is ahistorical. https://medium.com/@emilymenonbender/talking-about-a-schism-is-ahistorical-3c454a77220f. Accessed: 2025-05-13

Berryman, C.; Ferguson, C.J.; Negy, C. 2018. Social Media Use and Mental Health among Young Adults. *Psychiatric Quarterly* 89(2): 307-314. doi:10.1007/s11126-017-9535-6.

Bozzano, M. and Villafiorita, A. 2010. *Design and Safety Assessment of Critical Systems.* Boca Raton, FL: Auerbach Publications.

Brandt, M.J. and Proulx, T. 2016. Conceptual creep as a human (and scientific) goal. *Psychological Inquiry* 27(1): 18–23. doi:10.1080/1047840X.2016.1109577.

Burr, C and Leslie. D. 2023. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics* 3(1): 73–98. doi.org/10.1007/s43681-022-00178-0.

Cascardi, M. and Brown, C. 2016. Concept Creep or Meaningful Expansion? Response to Haslam. *Psychological Inquiry*, 27(1): 24–28. doi:10.1080/1047840X.2016.1111123.

Chandra, M.; Naik, S.; Ford, D.; Okoli, E.; De Choudhury, M.; Ershadi, M.; Ramos, G.; Hernandez, J.; Bhattacharjee, A.; Warreth, S.; Suh, J. 2025. From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). New Yotk: Association for Computing Machinery. Doi:10.1145/3715275.3732063.

Charmaz, K.; Harris, S.R.; Irvine, L. 2019. *The social self and everyday life: Understanding the world through symbolic interactionism.* Hoboken, NJ: John Wiley & Sons.

Chiodo, M.; Müller, D.; Siewert, P.; Wetherall, J. L.; Yasmine, Z.; Burden, J. 2025. Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility. arXiv:2505.10426.

Clymer, J.; Gabrieli, N.; Krueger, D; Larsen, T. 2024. Safety cases: How to justify the safety of advanced AI systems. arXiv:2403.10462.

Davies, M. and Birtwistle, M. 2023. Regulating AI in the UK, Technical Report. Ada Lovelace Institute.

Dobbe, R.I.J. 2022. System Safety and Artificial Intelligence. In The Oxford Handbook of AI Governance, edited by Bullock, J, B.; Chen, Y.; Himmerlreich, V, M.; Young, M, M.; Zhang, B, 441-458. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780197579329.013.67.

Furedi, F. 2016. The Cultural Underpinning of Concept Creep'. *Psychological Inquiry*, 27(1): 34-39. doi:10.1080/1047840X.2016.1111120.

Gabriel, S.; Puri, I.; Xu, X.; Malgaroli, M.; Ghassemi, M. 2024. Can AI relate: Testing large language model response for mental health support. arXiv:2405.12021.

Gyevnár, B. and Kasirzadeh, A. 2025. AI safety for everyone. *Nature Machine Intelligence* 7: 531–542. doi.org/10.1038/s42256-025-01020-y

Habli, I. 2025a. On the Meaning of AI Safety. In Proceedings of 20th European Dependable Computing Conference doi.10.1109/EDCC-C66476.2025.00055

Habli, I.; Hawkins, R.; Paterson.; Ryan, P.; Jia, Y.; Sujan, M.; McDermid, J. 2025b. The big argument for ai safety cases. arXiv:2503.11705.

Haidt, J. .2016. Why concepts creep to the left. *Psychological Inquiry* 27(1): 40–45. doi:10.1080/1047840X.2016.1115713.

Harding, J. and Kirk-Giannini, C.D. 2025. What is AI safety? What do we want it to be?. *Philosophical Studies* 182: 1495–1518. doi.org/10.1007/s11098-025-02367-z.

Harper, C.A., Purser, H. and Baguley, T. 2023. Do Concepts Creep to the Left and the Right? Evidence for Ideologically Salient Concept Breadth Judgments Across the Political Spectrum. *Social Psychological and Personality Science* 14(3): 319-332. doi:10.1177/19485506221104643.

Haslam, N. 2016. Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry* 27(1): 1–17. doi:10.1080/1047840X.2016.1082418.

Haslam, N.; Tse, J.S.Y.; De Deyne, S. 2021. Concept Creep and Psychiatrization. *Frontiers in Sociology* 6: 806147. doi:10.3389/fsoc.2021.806147.

Haslanger, S. 2023. Systemic and Structural Injustice: Is There a Difference?. *Philosophy* 98(1): 1–27. doi:10.1017/S0031819122000353.

Hendrycks, D.; Carlini, N.; Schulman, J.; Steinhardt, J. 2021. Unsolved problems in ml safety. arXiv:2109.13916.

Hendrycks, D. 2024. *Introduction to AI Safety, Ethics, and Society*. CRC Press. doi.org/10.1201/9781003530336

Hollnagel, E.; Wears, R.L.; Braithwaite, J. 2015. From Safety-I to Safety-II: a white paper. The Resilient Health Care Net: Published simultaneously by the University of Southern Denmark, University of Florida, USA, and Macquarie University, Australia.

IEC 61508. 2010. Functional safety of electrical/electronic/programmable electronic safety-related systems. The International Standard of the International Electrotechnical Commission.

IEC and ISO. 2015. Using and referencing IEC and ISO standards to support public policy. Geneva, Switzerland: International Electrotechnical Commission and International Organisation for Standardization.

ISO 26262. 2018. Road vehicles – functional safety. International Organisation for Standardisation.

Jia, Y.; McDermid, J.; Lawton, T.; Habli, I. 2022. The Role of Explainability in Assuring Safety of Machine Learning in Healthcare. *Transactions on Emerging Topics in Computing* 10(04): 1746-1760. doi:10.1109/TETC.2022.3171314.

Johnson, B. 2022. Metacognition for artificial intelligence system safety – An approach to safe and desired behavior. *Safety Science*, 151(105743) doi: 10.1016/j.ssci.2022.105743.

Kamoi, R. and Kobayashi, K. 2020. Out-of-distribution detection with likelihoods assigned by deep generative models using multimodal prior distributions. In Proceedings to CEUR Workshop Vol. 2560. CEUR-WS.

Knight, J.C. 2002. Safety Critical Systems: Challenges and Directions. In Proceedings of the 24th International Conference on Software Engineering (ICSE '02). New York: Association for Computing Machinery. doi:10.1145/581339.581406.

Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; Campos-Castillo, C. 2022. Too human and not human enough: A grounded

theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* 26(10): 5923-5941. https://doi.org/10.1177/14614448221142007

Lazar, S. and Nelson, A. 2023. AI Safety on Whose Terms?. *Science* 381(6653): 138. doi:10.1126/science.adi8982.

Leveson, N. 2023. *An Introduction to System Safety Engineering*. Cambridge, MA: MIT Press

Li, L.; Dong, B.; Wang, R.; Hu, X.; Zuo, W.; Lin, D.; Qiao, Y.; Shao, J. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. arXiv:2402.05044.

Lock, S. 2022. What is AI chatbot phenomenon ChatGPT and could it replace humans? The Guardian. http://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans. Accessed 2025-08-01

Lowrance. W. 1976. *Of acceptable risk: Science and the determination of safety*. Los Altos, Calif: W. Kaufmann.

Manuele, F.R. 2010. Acceptable Risk: Time for Professionals to Adopt the Concept. *Professional Safety* 55(6): 30–38.

Martinez, J.; Eguia, A.; Urretavizcaya, I.; Amparan, I.; Negro, P. L. 2023. Fault tree analysis and failure modes and effects analysis for systems with artificial intelligence: A mapping study. In Proceedings of 7th International Conference on System Reliability and Safety (ICSRS). IEEE. doi.10.1109/ICSRS59833.2023.10381456

Montag, C.; Yang, H.; Elhai, J.D. 2021. On the Psychology of TikTok Use: A First Glimpse From Empirical Findings. *Frontiers in Public Health*, 9: 641673. doi.10.3389/fpubh.2021.641673.

National Research Council. 1997. Digital Instrumentation and Control Systems to Nuclear Power Plant: safety and reliability issues. Washington DC: National Academic Press.

Niemi, L. and Young, L. 2016. Justice and the moral lexicon. *Psychological Inquiry* 27(1): 50–54. doi:10.1080/1047840X.2016.1111122.

Osman, M. 2025. Psychological harm: what is it and how does it apply to consumer products with internet connectivity?. *Journal of Risk Research* 28(2): 1-22. doi.org/10.1080/13669877.2025.2491086.

Pinker, S. 2011. *The Better Angels of our Nature*. London: Penguin.

Porter, Z.; Habli, I.; McDermid, J.; Kaas, M. 2024. A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI and Ethics* 4(2): 593-616. doi.org/10.1007/s43681-023-00297-2.

Qiu, J.; He, Y.; Juan, X.; Wang, Y.; Liu, Y.; Yao, Z.; Wu, Y.; Jiang, X.; Yang, L.; Wang, M. 2025. Emoagent: Assessing and safeguarding human-ai interaction for mental health safety. arXiv:2504.09689

Raji, I.D.; and Dobbe, R. 2023. Concrete problems in AI safety, revisited. arXiv:2401.10899.

Ren, R.; Basart, S.; Khoja, A.; Gatti, A.; Phan, L.; Yin, X;, Mazeika, M.; Pan, A.; Mukobi, G.; Kim, R.; Fitz, S.; D, Hendrycks. 2025. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS '24), Vol. 37. New York: Curran Associates Inc.

Rismani, S.; Shelby, R.; Smart, A.; Santos, R, D.; Moon, A.; Rostamzadeh, N. 2023. Beyond the ML Model: Applying Safety Engineering Frameworks to Text-to-Image Development. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). New York: Association for Computing Machinery. doi.org/10.1145/3600211.3604685

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.M.; Gallegos, J.; Smart, A.; Garcia, E.; Virk, G. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. Canada: Association for Computing Machinery. doi: 10.1145/3600211.3604673.

Stilgoe, J.; Owen, R.; Macnaghten, P. 2017. Developing a framework for responsible innovation. In The ethics of nanotechnology, geoengineering, and clean energy, edited by Maynard, M and Stilgoe, J, 347-359.London: Routledge.

Sujan, M.; Habli, I.; Kelly, T.; Pozzi, S.; and 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. doi.org/10.1016/j.ssci.2015.12.021

Swuste, P.H.J.J.; van Gulijk, C.; Zwaard, A.W. 2011. Accident causation and prevention, the start of Safety Science. In Loss Prevention and Safety, a practical risk management handbook, edited by Shariari, M, 1-18. Sweden: Chalmers University of Technology, Dept. of Product and Production Development.

Twenge, J.M. and Campbell, W.K. 2018. Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Preventive Medicine Reports* 12: 271-283. doi:10.1016/j.pmedr.2018.10.003.

UK Government. 2024. Frontier AI: Capabilities and Risks Discussion Paper. https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper. Accessed: 2025-08-01.

UK Department for Science, Innovation and Technology. 2023. Introducing the AI Safety Institute. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute. Accessed: 2025-08-01.

University of Oxford. 2023. Expert Comment: Oxford AI Experts Comment on the Outcomes of the UK AI Safety Summit. https://www.ox.ac.uk/news/2023-11-03-expert-comment-oxford-ai-experts-comment-outcomes-uk-ai-safety-summit Accessed: 2025-08-05.

Vylomova, E.; Murphy, S.; Haslam, N. 2019. 'Evaluation of semantic change of harm-related concepts in psychology. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. Italy: Association for Computational Linguistics. doi: 10.18653/v1/W19-4704.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C. 2022. Taxonomy of Risks posed by Language Models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L.A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I ,2023. Sociotechnical safety evaluation of generative ai systems. arXiv:2310.11986.

Wickens, C.D.; Helton, W.S.; Hollands, J.G.; Banbury, S. 2021. *Engineering psychology and human performance*. 5th Edn. New York: Routledge.