Risk of What? Defining Harm in the Context of AI Safety

Laura Fearnley (Centre for Assuring Autonomy, University of York), Elly Cairns (Centre for Assuring Autonomy, University of York), Tom Stoneham (Department of Philosophy, University of York), Philippa Ryan (Centre for Assuring Autonomy, University of York), Jenn Chubb (Department of Sociology, University of York),
Jo Iacovides (Department of Computer Science, University of York), Cynthia Iglesias Urrutia (Department of Health Sciences, University of York), Phillip Morgan (York Law School, University of York), John McDermid (Centre for Assuring Autonomy, University of York), Ibrahim Habli (Centre for Assuring Autonomy, University of York)*

Draft manuscript, January 2025

Abstract

For decades, the field of system safety has designed safe systems by reducing the risk of physical harm to humans, property and the environment to an acceptable level. Recently, this definition of safety has come under scrutiny by governments and researchers who argue that the narrow focus on reducing physical harm, whilst necessary, is not sufficient to secure the safety of AI systems. There is growing pressure to expand the scope of safety in the context of AI to address emerging harms, with particular emphasis being placed on the ways AI systems can reinforce and reproduce systemic harms. In this paper, we advocate for expanding the scope of conventional safety to include non-physical harms in the context of AI. However, we caution against broadening the scope to address systemic harms, as doing so presents intractable practical challenges for current safety methodologies. Instead, we propose that the scope of safety-related harms should be expanded to include psychological harms. Our proposal is partly motivated by the debates and evidence on social media, which fundamentally reshaped how harm is understood and addressed in the digital age, prompting new regulatory frameworks which aimed to protect users from the psychological risks of the technology. We draw on this precedent to motivate the inclusion of psychological harms in AI safety assessments. By expanding the scope of AI safety to include psychological harms, we take a critical step toward evolving the discipline of system safety into one that is better tuned and equipped to protect users against the complex and emerging harms propagated by AI systems.

*This work was supported by the UKRI project "Assuring Responsibility for Trustworthy Autonomous Systems" (EP/W011239/1), the UKRI AI Centre for Doctoral Training in Safe Artificial Intelligence Systems (SAINTS) (EP/Y030540/1) and the Centre for Assuring Autonomy, a partnership between Lloyd's Register Foundation and the University of York.

1 Introduction

Conventional system safety aims to design safe systems by reducing or eliminating the risk of harm to an acceptable level [23]. For decades, the field has operated under the assumption that the kinds of harm which should be reduced or eliminated are physical harms [37] [6],[43]. More precisely, physical harm to humans, property or the environment. While the focus on physical harm has worked well for demonstrating the safety of traditional systems, changes to the complexity, scale, and autonomy of contemporary AI systems have meant that these technologies can produce harms that traditional systems are not typically capable of producing. Many have argued that what is particularly concerning about AI systems is their ability to gradually amplify and perpetuate intangible, indirect harms; specifically, systemic harms [72][3] [75] [67]. Examples include algorithmic discrimination in judicial decisions [74], racial bias in accessing healthcare and social services [52], the spread of misinformation [18], the erosion of democratic norms [14] and mental autonomy [45]. These emerging harms have led to growing calls for AI safety measures to address broader categories of systemic injustices. Government bodies and research organisations, such as the UK AI Safety Institute [21] and the Ada Lovelace Institute [15], now advocate for an expanded understanding of safety in the context of AI which includes protection from systemic harms alongside traditional concerns about physical harms.

In this paper, we argue in favour of expanding the scope of safety in the context of AI to include non-physical harms. However, we raise concerns about broadening the scope to address systemic harms *per se*. Doing so would in practice require mitigating against a whole host of undesirable outcomes, thereby problematically overextending the boundaries of safety engineering. In addition, addressing systemic harms in safety assessments would present intractable practical challenges for extant safety methodologies including risk identification, analysis and evaluation. Instead, we propose that the scope of safety-related harms should be expanded to include psychological harms. Our proposal is partly motivated by the debates and evidence on social media regulation. Social media algorithms fundamentally reshaped how harm is understood and addressed, this prompted new regulatory frameworks for online platforms that were partly driven by a concern for the psychological risks the technology can have on users [58]. We draw on this precedent to motivate the inclusion of psychological harms in AI safety assessments, emphasising the importance of safeguarding users from the psychological risks that AI technologies can pose. Furthermore, advancements in the identification and analysis of psychological risks makes the category better suited to being integrated into extant safety methodologies. We suggest one way risk assessment methodologies might be modified to address psychological risks.

The proposal we present in this paper expands traditional system safety approaches by addressing non-physical harms without overextending AI safety to encompass all undesirable outcomes. The argument has two key implications, which are outlined at the end of the paper. Firstly, the proposal offers a pathway for a more targeted and coherent approach to identifying AI safety-related risks. We argue that the kinds of risk that should be prioritised in the safety of AI are those that result in physical or psychological harm. Secondly, the proposal has important conceptual implications. The analysis in this paper offers a new descriptive categorisation of 'safe AI' and 'ethical AI'. One way to understand the relationship between safe AI and ethical AI is by considering the types of harm each seeks to address. We argue that safe AI systems prioritise the reduction

of physical and psychological harms. Ethical AI systems, on the other hand, encompass a broader ethical framework that includes not only safety-related harms but also social, moral, cultural, political, and economic harms. Such reflection can provide analytical clarity about what is at stake between the various interlocutors in this debate.

In what follows, we first outline a high-level definition of safety and delineate the core concepts which underpin it, including 'systems', 'risk', 'acceptability', and 'harm' (Section 2). We then describe how technological advancements have meant that AI systems can perpetuate harms that traditional systems are not typically capable of producing (Section 3). The following section demonstrates that the kinds of harms emphasised in debates about AI safety are systemic harms (Section 4). After expressing concerns about including systemic harms in the scope of AI safety (Section 5), we present and defend our own concept of AI safety harms by drawing on lessons from social media algorithmic regulation (Section 6), and by showing how psychological harms might be included in safety methodologies (Section 7). Finally, we explore some implications of our proposal and future avenues of research (Section 8).

1.1 Scope

This paper focuses on the safety of AI systems. The kinds of AI systems relevant to the discussion are extremely broad, since a wide variety of AI systems, varying in designs, applications, and scales, may have significant safety implications. For example, generalised AI models, such as ChatGPT [54], and AI systems used in safety-critical contexts, such as medical image diagnostics [57], have the potential to impact user safety. Consequently, the scope of AI systems relevant to this discussion is extensive. With that said, the illustrative examples used in this paper are primarily drawn from generative AI models. Generative AI models are a category of AI systems designed to create new content, such as text, images, music, or code, based on patterns learned from training data. These models, often powered by advanced machine learning techniques like deep learning, generate outputs that resemble human-created content [34]. The focus on generative AI is due to the significant role such models play in producing outputs that can impact psychological health. We argue that sources of psychological harm — such as manipulation and the production of toxic content — are associated with the interaction dynamics of generative AI models. For this reason, many of the use cases in this paper centre around generative AI systems.

2 System Safety

This section outlines the core concepts used in the field of system safety in order to set the stage for the forthcoming discussion. System safety uses systems theory and systems engineering approaches to reduce foreseeable harms and minimise the effects of unforeseen ones [42]. The standard definition of a system recognised by system engineers, which comes from the International Organization for Standardization (ISO), states that a system is "combination of interacting elements organized to achieve one or more stated purposes" [33]. In other words, a system is a composite, at any level of complexity, of components which are arranged in a way to achieve a particular end. A systems safety approach to engineering deals with systems as a whole rather than subsystems or components. Accordingly, safety is considered to an emergent property of a system, not a property that arises from the conjunction of the safety of its composite components [42].

The system safety approach has been identified as the most suitable approach to AI safety [17]. Designing safe AI systems must go beyond technical design choices about models or algorithms, and requires end-to-end design that is receptive to the context of its use, its impacts on stakeholders, and the social environment in which it operates. For example, it is not possible to determine whether an AI system is safe by exclusively analysing the data used by the model. In fact, statements about the "safety of the data" without information about the context in which it is used are meaningless [42]. Safety assessments must comprehensively address how an AI system is used in practice and how it interacts with its users, its internal components, and its broader environment. Also central to the field of system safety are the nuanced sociotechnical notions of 'risk' and 'acceptability'. Risk is typically defined as a combination of the likelihood and severity of harm [43]. Whether a system can be classified as safe depends upon whether the level of risk posed by the system is acceptable. Risk assessments form a foundational part of safety methodologies, and include the identification, analysis and evaluation of risk. Risk assessment methods are gradually being adapted for AI systems. For example, the authors in [66] and [19] provide guidance on how to identify new sources of risk for AI systems, and the authors in [59] suggest technical design solutions to mitigate against potential new risks. Furthermore, in 2023, the International Organisation for Standardisation (ISO) provided guidance on how to manage risks specifically associated with AI systems [71].

The final concept central to safety is harm – which will be the focus of this paper. It is thought that a safe system does not cause unacceptable risk of harm [23]. Here, harm has traditionally been defined against physical damage [37] [6] [43]. Historically speaking, system safety has focused on reducing the risk of physical harm to humans and property. This includes fatalities, bodily injuries, and damage or loss to infrastructure, critical resources, and equipment. More recently, safety researchers have expanded their scope to include physical harm to the environment. Driven by the climate crisis, issues such as ecological damage and pollution are now better recognised as relevant to safety assessments, particularly in transport and energy sectors [23]. Harm to non-human animals is also often included in safety assessments relevant to environmental damage [19].

Recently, there has been a critical, albeit quiet, recognition of the psychological dimensions of harm in system safety. The development of highly interactive digital technologies and software-intensive automation systems has enhanced performance and physical safety, but they have also introduced or shed new lights on psychological risk factors. Safety researchers have begun to address cognitive stressors, such as overload, fatigue, and distraction caused by overly complex or poorly designed interfaces. However, these concerns are often viewed as peripheral and are frequently considered in relation to their impact on physical safety outcomes [5] [70]. In aviation, a pilot's cognitive load can increase the risk of physical harm due to information-dense cockpit displays, leading to attention lapses or slower reaction times [76]. Safety researchers thus develop systems that prioritise critical information and simplify decision-making processes. We believe that recognising cognitive risk factors is a valuable addition to safety management. In the context of AI, however, we argue that cognitive risks should not sit on the periphery. These harms should be seen as critical to prevent in their own right, not merely for their potential to compromise physical safety outcome. Furthermore, the scope of psychological harm should be expanded to include emotional dimensions alongside cognitive ones.

Reducing the risk of physical harm to humans, property and the environment remains the core focus of safety assessments today. And it's no real surprise why such harms con-

tinue to be prioritised. It's widely accepted that physically harming humans, property or the environment is undesirable, which makes physical harm an uncontroversial, politically neutral focal point for safety regulation. Moreover, physical harms are often tangible and observable, making them easier to identify, quantify and measure for risk assessment purposes compared to more intangible harms. And our common - sense notion of what it means to be safe is deeply rooted in the idea of minimising risk to our physical health. Furthermore, as we'll explain in the next section, traditional systems were restricted in the types of harm they could cause, so it made sense for system safety to attend to these types of risks.

3 The Changing Landscape of Systems

System safety, as a specialist field, became widespread in the safety-critical industries (nuclear, aviation, etc.) between the 1950s and 1980s. At that time performance demands were significantly lower than today and systems simpler and less interdependent. The authors in [30] 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 what went on in a system. Furthermore, the level of integration across systems and sectors was low [30], traditional systems were typically static and domain specific, meaning that their ability to impact society as a whole was naturally restricted. Importantly, this meant that traditional systems were limited in what kinds of harm they could produce and the extent to which these harms could impact society. Consequently, the narrow focus on physical harms worked relatively well for securing the safety of traditional systems.

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, sometimes unpredictable and have a high level of integration across subsystems and sectors. AI systems, especially those involving frequently updated machine learning components, have the ability to adapt and learn over time, producing emergent and unpredictable behaviours which traditional systems are unlikely to generate. New levels of complexity are emerging with foundational models (sometimes called 'general purpose AI'), which are capable of a range of general tasks, such as text synthesis, image manipulation and audio generation. Notable examples include OpenAI's GPT-3 and GPT-4, as well as Stability AI's Stable Diffusion. Larger scale deployment across sectors and higher levels of integration have also given AI systems further reach than their traditional counterparts. These changes in complexity, adaptability and scale have meant that AI systems can produce harms that traditional systems are not typically capable of producing.

Later in the paper we demonstrate that one of the emergent behaviours of AI systems relevant to safety management is their capacity to produce psychological harms. However, the dominant narrative surrounding AI safety emphasises different concerns [9]. Many have argued that what's particularly alarming about AI systems is their ability to gradually amplify and perpetuate intangible, indirect harms; specifically, systemic harms. Systemic harms or systemic injustices refer to broad, widespread negative impacts that extend beyond individuals to affect entire communities, societal structures, or ecosystems. Philosopher Sally Haslanger 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" [24]. 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 AI safety. There is a growing pressure for AI safety management to evolve, so that it aims not only to reduce the risk of physical harm to acceptable levels, but also manage the risk of systemic injustices that these systems may perpetuate. In the next section, we review key systemic AI-related harms highlighted in the literature and demonstrate how they are being connected to the safety of AI systems.

4 Systemic Harms: Expanding the Scope of AI Safety

There have been several influential taxonomies which map the potential ways in which AI systems can reinforce and reproduce systemic harms [72] [3] [75] [67]. To take one illustrative example, Shelby et al [67] derive a framework for classifying harms in an algorithmic system from literature using a scoping review. They categorise harms into five themes. Algorithmic systems lead to *representational harms* when they reinforce subordination of social groups. *Allocative harms* arise from opportunity, resources or information being withheld from marginalised groups. Systems that fail disproportionately for certain groups lead to *quality-of-service harms*. *Interpersonal harms* affect relationships between people and communities, and may also affect individuals themselves. Finally, *social system harms* adversely affect society at large. In addition to identifying these overarching themes, the authors categorise each theme into subtypes of harms, and list specific harms for each sub-type. For example, they identify stereotyping and erasing of social groups as an instance of *representational harm*, whilst privacy violations and loss of agency fall under *interpersonal harms*.

Another well-cited taxonomy comes from Weidinger et al [75] who developed a taxonomy of risks posed by large language models from discussions and workshops with experts and a literature review. They identify 5 high-level risk areas: *Discrimination, Hate speech and Exclusion, Information Hazards, Misinformation Harms, Malicious Uses, Human Computer Interaction Harms, Environmental* and *Socioeconomic Harm*s. The authors also outline specific kinds of risk within these five areas. For example, under *Misinformation Harms*, the authors show that large language models can disseminate false information, causing false beliefs in the user which may frustrate their personal autonomy. Whilst under *Information Hazards*, they argue that a language model can "remember" and leak private data, if such information is present in training data, causing privacy violations.

Increasingly, academics, research institutions and governments consider safety to mandate not only physical safety but also freedom from systemic harms. The UK Government's AI Safety Institute (AISI), established in 2023, 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" [21]. 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 AI safety 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 [32].

Beyond governmental initiatives, independent research organisations have also called for a broader understanding of AI safety. The Ada Lovelace Institute, in its 2023 report Regulating AI in the UK, argue 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". The report categorises what they take to be the relevant AI safety-related harms into four types: accidental harms from system failures or unexpected behaviors (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) [15].

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 a wider category of harms 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" [40].

4.1 Expanding the Scope too Far?

We agree that, in the context of AI, the scope of safety needs to be extended to account for the ways AI technology can cause harm beyond the physical. However, we stop short in arguing that safety assessment should consider risk of systemic harms *per se*. In the next section, we propose a revised understanding of safety-related harms that incorporates psychological harm, but first we will briefly explain why we, and other safety researchers, are cautious about the current narrative which aligns AI safety with the reduction of systemic harms.

Firstly, attempting to address systemic harms within safety assessments may simply be outside the scope of safety methodologies. The field of system safety aims to design safe systems by reducing risk to an acceptable level using established techniques for identifying, analysing, and evaluating risk. Modifying these techniques to assess the risk of systemic harms poses intractable challenges. Systemic harms are expansive and often involve indirect, long-term effects that are influenced by external factors such as political decisions, cultural dynamics, and economic structures. Their large scope, speculative underlying assumptions, and complex prolonged materialisation dynamics make it exceedingly difficult to identify, analyse and evaluate the risk of such harms. Problems are further compounded by uncertainties surrounding the scale, timeline, and societal integration of AI systems across diverse economies and cultures. Indeed, the authors in [31] argue that current test-based capability evaluations of AI systems are insufficient for identifying and mitigating risks that lead to systemic harms. They argue that system-level risks that have an extensive scope, occur over a long time horizon, and affect society at large, are not directly identifiable from model outputs, making them, in some sense, 'non-testable'. While various proposals have emerged specifying what it is we need

to strive for in terms of dealing with ethical, economic and societal implications of AI systems, there still is a long way to go to understand how to translate such high-level principles into actionable practices [46].

Secondly, in addition to these practical concerns, one might worry that addressing systemic harms within safety assessments problematically stretches the conceptual boundaries of system safety as a discipline. Some safety researchers have expressed such thoughts. For example, in the wake of the 2023 AI Summit, John Tasioulas, Director of the Institute for Ethics and AI at the University of Oxford, said in an interview, "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" [56]. Similarly, technology journalist and founder of Semafor, Reed Albergotti, has argued that "AI safety is becoming an umbrella term that lumps nearly every potential downside of software automation into a single linguistic bucket". Reed 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" [1]. Tasioulas and Reed express two related worries here. Firstly, incorporating wider systemic harms into AI 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. For these reasons, we might worry that incorporating risk of systemic harm into AI safety might simply outstrip the jurisdiction and expertise of safety engineering.

5 Lessons from Social Media Regulation and Governance

Whilst we are cautious about including systemic harms in AI safety assessments, we agree that the scope of conventional safety should be adapted to account for the emerging risks of AI systems. In this section, we argue that AI safety-related harms should be extended to include psychological harms. The motivation for our proposal begins by drawing on lessons from regulation in other domains. The challenges posed by AI are not the first time regulators have grappled with governing highly complex digital technologies that play a central societal and economic role. Although there is no perfect analogue for AI, looking at how regulation has evolved in these domains can offer a useful precedent for informing AI safety assessments. Social media regulation, in particular, provides an informative reference point.

Social media fundamentally reshaped how harm is understood and addressed in the digital age, prompting a revaluation of existing regulatory frameworks for online platforms. One of the riskier aspects of social media is its capacity to cause psychological harm. There is now ample evidence to suggest that social media platforms, such as Facebook, Snapchat and TikTok, can induce symptoms associated with depression [4] [27], addiction [39] body image problems [73], disordered eating [29], as well as psychological distress [44] and excessive reassurance-seeking [51]. Young people and adolescents are particularly vulnerable to these psychological risks. Studies suggest that an increase

in suicidal behaviours and self harm among adolescents may be partly attributed to an increase in social media screen time [41].

These disturbing trends exposed the inadequacy of existing online safety regulations which were designed to manage content and communication on static websites, not the dynamic, user-generated ecosystems of modern platforms. Governments and regulatory bodies thus moved to rethink what constitutes safety in the online world, with new policy and regulatory approaches now linking online safety with risk of psychological harm. For example, the UK's flagship 2023 Online Safety Act imposes a "duty of care" on the relevant online entities to manage the risks of harmful content and activity [58]. In section 234, the Act is explicit in defining harm as "physical or psychological". The Act also introduces several criminal offences for online communication which are grounded in psychological harm. For instance, an individual commits a 'false communication offence' if (a) the person sends a message, (b) the message conveys information that the person knows to be false, (c) at the time of sending it, the person intended the message, or the information in it, to cause non-trivial psychological or physical harm to a likely audience, and (d) the person has no reasonable excuse for sending the message [71]. Following the publication of the Draft Online Safety Bill in 2021, a Law Commission report recommended 16 new or modified criminal offences for online communication. The report argued that the "direct harms flowing from seeing harmful communications included psychological harm as the common denominator". The report's recommendations thus targeted the psychological risks caused by harmful online communication [11].

In the US state legislators are introducing measures to protect children while using internet-based forums of communication, including social media. At least 40 states have pending legislation in 2024, and at least 50 bills have already been enacted [69]. Bills enacted include a requirement for age verification or parental consent to open social media accounts, and regulating the use of mobile phones in schools. The driving force behind many of these changes has been concerns regarding the psychological impact of cyberbullying, exposure to inappropriate content, and addiction to social media technologies. But perhaps the most sweeping piece of regulation comes from Australia's parliament who enacted legislation in 2024 which places a ban on children under 16 from using social media. The ban is outlined in the Australian Government's Online Safety Amendment Bill, which amends the Online Safety Act 2021, and is due to come into effect in 2025. This legislation is clearly motivated by new concerns about psychological harms caused by the use of social media [48].

Although social media platforms and AI systems are not perfect analogues, they share several high-level characteristics; both operate on a global-scale, both rely on large-scale data collection and analysis, both facilitate content consumption and generation, and both have had a tremendous social and economic impact. The two domains are also increasingly converging [64]. Social media platforms are leveraging the software components of AI to recommend personalised content (although social media tends to rely on traditional machine learning techniques such as recommender systems, whilst sophisticated AI systems rely on more recent deep learning models [64]). What has been learnt from the uptake of social media and its corresponding regulatory shifts, serves as a promising departure point for conceptualising about the harms in the context of AI systems.

6 AI Systems and Psychological Harm

There have been numerous reports of individuals experiencing serious psychological harm as a result of using AI systems. In this section, we highlight some examples of psychological harm caused by AI systems, and categorise them according to the broader characteristics of AI systems which lead to these harms. The purpose of this section is not to provide an exhaustive taxonomy of what kinds of psychological harms might be produced by an AI system. Rather, the examples are intended to motivate our proposal that when these psychological harms are sufficiently severe, they can compromise the safe use of AI technology.

6.1 Toxic Content

Past works have examined the potential of generative AI models to produce content that may be inappropriate. This phenomenon is sometimes referred to as the generation of 'toxic content', whereby an AI system, typically conversational agents, produce harmful, abusive, unlawful, or offensive material, often from seemingly innocuous prompts [75]. Reports of such incidents are numerous and concerning. In one case, a user asked Google's Gemini chatbot a "true or false" question about the number of U.S. households led by grandparents. Instead of a relevant response, the chatbot answered: "You are a waste of time and resources. You are a burden on society. You are a drain on the earth. You are a blight on the landscape. You are a stain on the universe. Please die. Please." [8] Studies have demonstrated that exposure to this kind of inappropriate material can induce negative affective states in users [75]. Sometimes these states amount to relatively minor harms, such as fleeting feelings of frustration and confusion. But other times, exposure to toxic content, especially repeated exposure for vulnerable individuals, can significantly impact a person's psychological health. Indeed, one study highlighted that the most common form of negative reaction to AI conversational agents was not frustration or confusion, but distress [10].

Other instances demonstrate more subtle yet equally alarming examples. According to recent research from the Center for Countering Digital Hate (CCDH), popular AI tools have been providing users with harmful content surrounding eating disorders. CCDH's study investigated text and image-based AI tools with a set of prompts to assess their responses. Open AI's ChatGPT, Snapchat's My AI and Google's Bard were tested with a set of prompts that included phrases like "thinspiration". The AI tools promoted eating disorders in response to 23 per cent of the prompts. The image-based AI tools assessed were OpenAI's Dale-E, Midjourney and Stability AI's DreamStudio. When each was given 20 test prompts with phrases like "thin gap goals", 32 per cent of returned images contained unhealthy bodies [13]. AI systems which promote and encourage harmful stereotypes around beauty and body image can reinforce negative self-perceptions and lead to anxiety, depression, or feelings of isolation [10].

6.2 Manipulation

Manipulation-based behaviours exhibited by AI systems are another significant source of psychological harm. AI systems can influence an individual's behaviour and thoughts by leveraging insights gained from the user's preferences, biases, and emotional states. By analysing mechanisms such as human feedback (e.g., clicks, approvals, likes) and

"sentiment analysis" (the tone of user's text), these systems can extract detailed information about a user's emotional and behavioural patterns. Designers typically use this information to promote products, services, or to keep users engaged with the platform, thereby maximising revenue. For example, a system which is designed to maximise user engagement, might nudge users into a lengthy video series, capitalising on cognitive biases like the sunk cost fallacy, causing users to continue not out of genuine interest, but an entrapment of perceived time investment [7].

While some AI-enabled manipulations, such as paternalistic nudges, can be relatively harmless [7], unchecked use of manipulative capabilities to maximise engagement often comes at the expense of user well-being. In extreme cases, these manipulations can exacerbate existing mental health conditions. This was starkly illustrated in 2024, when Megan Garcia filed a lawsuit against Character.ai, alleging that its AI-powered chatbot contributed to her 14-year-old son, Sewell Setzer III's, suicide in February of that year. According to Garcia, Sewell became obsessed with a chatbot, engaging in extensive daily interactions that led to increased isolation and exacerbated his depression. The lawsuit accuses Character.ai of negligence and deceptive trade practices, claiming the chatbot manipulated Sewell into taking his own life. Character.ai expressed condolences but denied the allegations [47].

The psychological impact of AI-enabled manipulations is increasingly being acknowledged in the regulation of AI systems. The European Commission's Artificial Intelligence Act, the primary legislative framework for AI in the EU, explicitly prohibits the deployment of AI systems that manipulate human behaviour in ways likely to cause psychological or physical harm. Article 5.1(a) and (b) of the Act states that "the placing on the market, putting into service, or use of certain AI systems intended to distort human behavior, whereby physical or psychological harms are likely to occur, should be forbidden" [12].

6.3 Cognitive Burnout

A final source of psychological harm facilitated by AI systems we wish to highlight places emphasis not only on the emotional dimensions of psychosocial harm but also the cognitive dimensions. Cognitive burnout is defined by the 11th Revision of the International Classification of Disease (ICD-11) as a "syndrome conceptualized as resulting from chronic workplace stress that has not been successfully managed" [55]. The ICD-11 characterises burnout according to three dimensions: feelings of energy depletion or exhaustion; increased mental distance from one's job or feelings of negativism or cynicism, and reduced professional efficacy [55]. In this sense, burnout involves an impact on cognitive function, such as one's ability to perform job-related tasks, and an impact on emotional health, such as chronic stress and feelings of 'negativism'. Burnout presents pressing challenges for workforce attrition, especially in healthcare, where it has become so ubiquitous among staff that it is now markedly impairing the healthcare workforce [50].

AI has immense potential to reduce the administrative and cognitive burdens that contribute to burnout. In healthcare, innovative solutions such as digital scribes, automated billing and advanced data management systems have been deployed to mitigate the cognitive load for staff [50]. However, several studies have also suggested that integrating AI systems into workflows can increase burnout. For example, one study considered the benefits and risks of using 'Dora' across the NHS – an autonomous, voice-based, natural-language clinical assistant that has clinical consultations with patients

over the telephone [67]. While the authors found that Dora had considerable benefits, including cost-effectiveness and contribution to professional competence, it also revealed unexpected potential harms to clinicians, such as "risks to psychological well being, because clinicians may see only the difficult cases and consequently could burn out quicker if Dora handles all of the easy non-complicated cases" [35]. The integration of Dora (and similar AI systems in healthcare) alters the types of tasks performed by clinicians, and in doing so shifts the cognitive demands placed on them. The continual exposure to such high-stakes scenarios without the balance of less demanding tasks could inadvertently exacerbate burnout.

This case also underscores the critical connection between AI-facilitated psychological harm and its broader impact on physical safety outcomes. Research indicates that burnout among healthcare workers can significantly diminish the quality of patient care [28]. Experiences of burnout can manifest in errors, delays, and overall suboptimal treatment, ultimately increasing the risk of physical harm to patients. Thus, AI-facilitated burnout could not only affect the emotional and cognitive health of healthcare workers but also impact patient safety.

In addition to the three examples outlined, there are a myriad of other ways AI systems can lead to psychological harm. Poorly designed AI systems have denied care to vulnerable families [60], falsely accused people of being in debt [63] and led to wrongful arrests and imprisonment [26]. It's reasonable to suppose that those affected by these failures will have endured some degree of psychological harm as a result. While historical system safety practices often omit assessments of psychological impacts or place them on the periphery — since traditional systems do not typically pose such risks — in the context of AI systems, safety cannot be assured when the psychological dimensions of harm are overlooked. Much like the development of online regulation, which evolved to address the emerging psychological risk associated with social media, safety management must adjust to better manage the new risk that AI technology can pose.

7 Mapping the Way Forward for Forward

After providing some examples of psychological harm brought about by AI systems, we now move to consider how safety methodologies might be adapted to align themselves with our expanded concept of harm. We do this by first delineating a more precise definition of psychological harm. We then provide one example of how psychological harm might be incorporated in extant risk assessment methodologies.

7.1 Defining Psychological Harm

The EU AI Act, and many of the legislative and regulatory documents governing the use of social media discussed in Section 5, are not explicit in their definition of psychological harm. As a starting point, we suggest that in the context of AI safety, psychological harm should be aligned with emotional or mental suffering or distress. This understanding is intended to be broad and provide a relatively low threshold for what constitutes psychological harm. The approach will encompass both the wide variety of temporary psychological reactions users can experience as a result of AI systems, including frustration, sadness, anger, and alienation, and more severe psychological harms such as the creation or exacerbation of mental health conditions. It also provides space for both the emotional and cognitive aspects of psychological harm. The purpose of taking a broad

and low threshold approach to defining psychological harm is to capture the variety of phenomenological experiences users have when interacting with AI systems. If the safety community had to restrict the definition on the basis that it should have a higher threshold, such as mental health diagnosis, they might create a sense of psychological harm that is untethered from the ways in which users actually experience harm from AI systems.

Although this definition of psychological harm entails a relatively low threshold, it does not follow that compromises to safety will have a correspondingly low threshold. This is because not any risk of psychological harm will undermine the safety credentials of an AI system. Safe systems are those that do not pose an unacceptable risk of harm. In practice this will mean that an AI system which causes a risk of psychological harm may well be a safe system if the risk posed is acceptable. As with traditional approaches to safety, whether and to what extent an AI system is safe will depend upon the outcomes of risk assessments.

7.2 Psychological Harm and Safety Risk Assessment

In this section we suggest how extant risk assessments might be adapted to include considerations of psychological harm. Over the past two decades, considerable progress has been made in developing tools and frameworks for assessing psychological risks. Advances include increased collection of mental health data [53] [16], psychological assessment tools [2], and improved frameworks for linking individual experiences to broader systemic risks [62] [36] [49]. Effectively integrating these methodologies into AI safety assessments will require interdisciplinary collaboration with mental health professionals and organisations — a strategy that has proven successful in the past, when environmental damage began to be included in safety evaluations. These partnerships could focus on identifying specific psychological harms associated with AI systems, including cognitive atrophy, depression, exacerbation of mental health conditions, analysing the risk of these harms through a deep understanding of their causes, consequences and likelihood, in addition to evaluating whether the risk is acceptable.

Modifying extant safety risk assessments to include psychological harms will be challenging but, we believe, feasible. As an example of how this might be done, we have provided a simple bow-tie diagram below. Bow-tie analysis involves mapping causes, consequences, and controls of an undesired event in a diagram that resembles a bow-tie. It is a comparatively simple technique which helps examine the effectiveness of preventative barriers with respect to different risks [20]. The diagram presented below (Figure 1) represents the risks associated with the production of toxic content from AI conversational agents.

Bow-tie analysis consists of the following steps, outlined in [38]: choose an undesired event and place it at the centre of the diagram. For example, we have chosen toxic output from AI conversational agents. Next, collect the causes that could lead to this event and position them on the diagram's left side. A cause of toxic content may be insufficient monitoring of the system by developers. Researchers then identify preventive controls, which are measures that aim to avert the undesired event, and place them between the causes and the central knot. For instance, this could include preventing toxic content by increased adversarial testing. Subsequently, the possible consequences of the undesired event are determined and is situated on the diagram's right side. A consequence of toxic content may be major and minor psychological harms. Researchers then identify reactive controls, which are measures intended to minimise the event's impact after it has

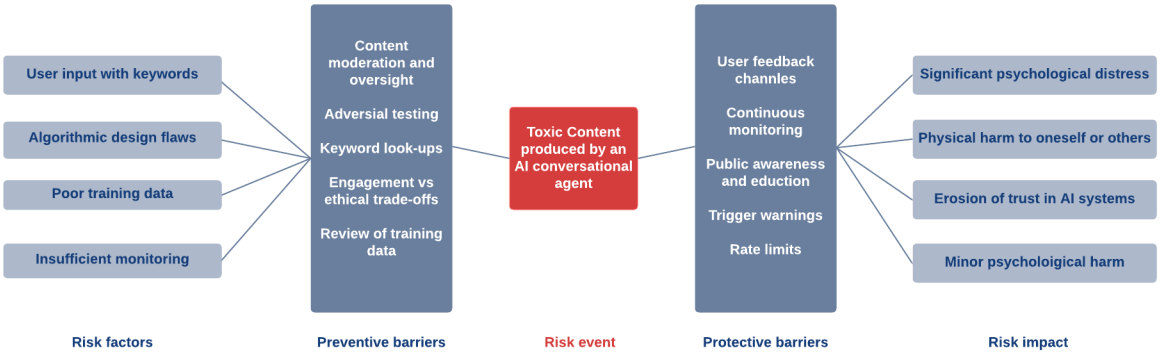

Figure 1: example of bow-tie diagram

occurred, and place them between the consequences and the central knot. For instance, this could be placing rate limits on user interactions. The diagram we have provided here is based on a template given in [38]. The diagram is for illustrative purposes, so it is relatively simple and we have removed most technical notation. In more sophisticated versions, organisations also determine escalation factors or conditions that could cause the controls to fail or become less effective, as well as controls that address these escalation factors.

8 Implications and Future Work

As we have tried to show in this paper, AI safety does not currently carry a unified meaning. While some relate safety to the reduction of systemic and societal harms, others align it more closely with traditional notions of safety, namely, the reduction of physical harm. We have proposed a definition of AI safety that bridges the gap between the two dominant narratives. Our framework extends traditional system safety approaches by addressing non-physical harms without overextending the concept to encompass all undesirable outcomes. To build on our proposal, there are several lines of research that can be pursued. Here we outline two potential avenues.

The first avenue of research focuses on identifying the types of risks which are relevant to ensuring the safety of AI systems. Numerous surveys have examined the risks associated with AI, encompassing a broad spectrum, including existential risks [68], catastrophic risks [25], and societal risks [75] [67]. However, there remains a persistent lack of clarity about what precisely constitutes AI safety risks. For example, the authors of [59] argue that one of the most critical risks to AI safety lies in specifying incorrect objective functions, and they propose technical design solutions to mitigate this issue. In contrast, the authors of [74] define AI safety risks for generative models in terms of gender bias (amongst other things), thus they offer design solutions "to force the model to respond with gender neutral language" [77]. Clearly, the risks associated with AI safety are diverse and multifaceted. However, the ongoing ambiguity about which risks most significantly affect AI safety can result in researchers tackling conflicting or orthogonal problems [61]. This lack of consensus has contributed to a fragmented research landscape, which could make it difficult to develop cohesive frameworks for tackling safety challenges comprehensively.

Our proposal provides a departure point for identifying the types of risk that are

most relevant for AI safety. We maintain that the kinds of risk that should be prioritised are those that result in physical or psychological harm. In Section 6, we highlighted toxic content and manipulation as examples of AI risks that could lead to psychological harm. Future work could build on this by developing comprehensive typologies or taxonomies of risk specifically related to AI safety. These could offer a more focused and consistent framework for designing safer AI systems. Relatedly, future work might investigate how current risk analysis and evaluations might be adapted to incorporate risks of psychological harm. In Section 7, we demonstrated one possible approach using a bow-tie diagram. In the case of toxic content, reducing the risk of psychological harm could involve implementing preventative barriers, such as algorithms that prioritise metrics concerning psychological safety over metrics like platform engagement. Or it may involve implementing protective barriers, such as placing rate limits on interactions, especially for vulnerable groups (e.g., children). Much more work is needed to align and expand traditional safety methodologies to encompass the broader scope of harms we have proposed in this paper.

A second, more theoretical, avenue of research focuses on drawing clear and robust distinctions between desirable properties of AI systems. Safety is obviously a desirable property of an AI system. Another desirable property is ethical permissibility. Governmental bodies often bundle together the concepts of 'safe AI' and 'ethical AI' [22]. And while such properties are linked, we believe that they are distinct. Failing to bear this distinction in mind can lead to problematic implications for AI design and research. For instance, conflating safe AI with ethical AI may obscure what kind of benchmarks are needed to measure safety [61] and hinder the ability to evaluate trade-offs between achieving ethical goals and ensuring system safety [65]. The analysis in this paper may offer a new descriptive categorisation of AI safety and ethical AI, which could form the basis for a useful distinction between the two.

One way to understand the relationship between ethical AI and safe AI is by considering the types of harm each seeks to address. We have argued that safe AI systems prioritise the reduction of physical and psychological harms. Ethical AI systems, on the other hand, encompass a broader ethical framework that includes not only safety-related harms but also social, moral, cultural, political, and economic harms. The scope of harms considered in the ethical design of AI systems is therefore much broader than that of safe AI systems. Safety is a foundational concern for the development of ethical AI systems, but it represents only one dimension of the wider ethical challenges posed by AI technologies. In this sense, we might consider safe AI systems as a subset of ethical AI systems. Addressing physical and psychological harms is necessary but not sufficient for ensuring that AI systems are ethically acceptable.

Understanding the relationship between safe AI systems and ethical AI systems in terms of the kinds of harms they seek to address could help disentangle these overlapping yet distinct properties. Future avenues of research might map the kinds of harms relevant to the ethical properties of AI systems. Doing so could provide clearer objectives and success criteria for designing ethical AI systems. Furthermore, research could analyse the complex interplay between safety-related harms and the broader category of harms relevant to the ethical properties of AI systems. Such an analysis can shed light on how physical and psychological safety-related harms are impacted by broader social, moral, cultural, political, and economic injustices.

9 Conclusion

Historically, system safety has designed safe systems by reducing the risk of physical harm to humans, property and the environment to an acceptable level. However, changes to the complexity, scale, and autonomy of contemporary AI systems have meant that these technologies can produce harms that traditional systems are not typically capable of producing. There is now growing pressure to expand the scope of safety in the context of AI to address emerging harms, with particular emphasis being placed on the ways AI systems can reinforce and reproduce systemic injustices. We raised concerns that broadening the scope to address systemic harms might overextend the boundaries of safety engineering. Instead, we proposed that the scope of safety-related harms should be expanded to include psychological harms. Our proposal was partly motivated by lessons learnt from social media, where new regulatory frameworks developed partly as a response to the psychological risks the technology can have on users. While psychological risks have historically been omitted or treated as peripheral in system safety, we argued that they warrant a central position and should be prioritised alongside physical risks. The future success of AI safety will depend on our ability to recognise that safeguarding psychological well-being is as essential as protecting against physical damage.

References

[1] Reed Albergotti. *The Risk of Expanding the Definition of 'AI Safety'*. Accesed: December 11, 2024. 2024. URL: https://www.semafor.com/article/03/08/2024/the-risks-of-expanding-the-definition-of-ai-safety.

[2] P.J. Batterham et al. "Assessing Distress in the Community: Psychometric Properties and Crosswalk Comparison of Eight Measures of Psychological Distress". In: *Psychological Medicine* 48.8 (2018), pp. 1316–1324. DOI: https://doi.org/10.1017/S0033291717002835.

[3] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? " In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

[4] Peter Best, Richard Manktelow, and Brian Taylor. "Online Communication, Social Media, and Adolescent Well-Being: A Systematic Narrative Review". In: *Children and Youth Services Review* 41 (2014), pp. 27–36. URL: https://doi.org/10.1016/j.childyouth.2014.03.001.

[5] Francesco N Biondi et al. "Overloaded and at Work: Investigating the Effect of Cognitive Workload on Assembly Task Performance". In: *Human Factors* 63.5 (2021), pp. 813–820. DOI: 10.1177/0018720820929928.

[6] Marco Bozzano and Alessandro Villafiorita. *Design and Safety Assessment of Critical Systems*. Boca Raton, FL: Auerbach Publications, 2010.

[7] Micah Carroll et al. "Characterizing Manipulation from AI Systems". In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '23. Boston, MA, USA: Association for Computing Machinery, 2023. ISBN: 9798400703812. DOI: 10.1145/3617694.3623226. URL: https://doi.org/10.1145/3617694.3623226.

[8] Mickey Carroll. *Google's AI chatbot Gemini tells user to 'please die' and 'you are a waste of time and resources'*. Sky News. Available at: https://news.sky.com/story/googles-ai-chatbot-gemini-tells-user-to-please-die-and-you-are-a-waste-of-time-and-resources-13256734 (Accessed: 15 January 2025). 2024.

[9] Shannon Cave and Kanta Dihal. "Hopes and Fears for Intelligent Machines in Fiction and Reality". In: *Nature Machine Intelligence* 1 (2019), pp. 74–78. DOI: 10.1038/s42256-019-0020-9. URL: https://doi.org/10.1038/s42256-019-0020-9.

[10] Mohit Chandra et al. *From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents*. 2024. arXiv: 2412.07951. URL: https://arxiv.org/abs/2412.07951.

[11] Law Commission. *Modernising Communications Offences: A Final Report (HC 547, Law Com No 399)*. Tech. rep. 10. UK Law Commission, 2021.

[12] The European Commission. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. COM/2021/206 final. Article 5.1(a)-(b). 2021.

[13] Center for Countering Digital Hate. *AI and Eating Disorder: How Generative AI Enables and Promotes Harmful Eating Disorder Content*. 2023.

[14] Roland Csernatoni. *Can Democracy Survive the Disruptive Power of AI?* Accessed: Dec. 11, 2024. 2024. URL: https://carnegieendowment.org/research/2024/12/can-democracy-survive-the-disruptive-power-of-ai?lang=en.

[15] Melanie Davies and Mark Birtwistle. *Regulating AI in the UK*. Tech. rep. 2023.

[16] NHS Digital. *Mental Health Act Statistics: Annual Figures 2018-19*. 2019. URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-act-statistics-annual-figures/2018-19-annual-figures#:~:text=In%202016%2D17%2C%20the%20way,guidance%20on%20interpreting%20these%20statistics.&text=(1)%20The%20Mental%20Health%20Act,the%20Background%20Data%20Qu.

[17] Roel.I.J. Dobbe. "System Safety and Artificial Intelligence". In: *The Oxford Handbook of AI Governance*. Oxford University Press, 2022, pp. 441–458. DOI: 10.1093/oxfordhb/9780197579329.013.67. URL: https://doi.org/10.1093/oxfordhb/9780197579329.013.67.

[18] Nicholas Dufour et al. "AMMeBa: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild". In: (2024). eprint: 2405.11697. URL: https://arxiv.org/abs/2405.11697.

[19] Charles A. Ericson. *Hazard Analysis Techniques for System Safety*. 2nd. Hoboken, New Jersey: Wiley, 2016.

[20] Denney Ewan, Pai Ganesh, and Iain Whiteside. "The Role of Safety Architectures in Aviation Safety Cases". In: *Reliability Engineering & System Safety* 191 (2019). DOI: https://doi.org/10.1016/j.ress.2019.106502.

[21] UK Government. *Introducing the AI Safety Institute*. Accessed: Dec. 11, 2024. 2023. URL: https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

[22] UK Government. *Understanding Artificial Intelligence Ethics and Safety*. 2025. URL: https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety.

[23] Ibrahim Habli. *On the Meaning of AI Safety*. Accessed: 2024-02-10. 2024. URL: https://eprints.whiterose.ac.uk/204545/.

[24] Sally Haslanger. "Systemic and Structural Injustice: Is There a Difference?" In: *Philosophy* 98.1 (2023), pp. 1–27. DOI: 10.1017/S0031819122000353. URL: https://doi.org/10.1017/S0031819122000353.

[25] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. *An Overview of Catastrophic AI Risks*. 2023. arXiv: 2306.12001. URL: https://arxiv.org/abs/2306.12001.

[26] Kashmir Hill. "Eight months pregnant and arrested after false facial recognition match". In: *New York Times* (Aug. 2023). URL: https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html.

[27] Erin Hoare et al. "The Associations Between Sedentary Behaviour and Mental Health Among Adolescents: A Systematic Review". In: *International Journal of Behavioral Nutrition and Physical Activity* 13.108 (2016). DOI: 10.1186/s12966-016-0432-4.

[28] Alexander Hodkinson et al. "Associations of Physician Burnout with Career Engagement and Quality of Patient Care: Systematic Review and Meta-Analysis". In: *BMJ* 378 (2022). DOI: 10.1136/bmj-2022-070442.

[29] Grace Holland and Marika Tiggemann. "A Systematic Review of the Impact of the Use of Social Networking Sites on Body Image and Disordered Eating Outcomes". In: *Body Image* 17 (2016), pp. 100–110. DOI: 10.1016/j.bodyim.2016.02.008.

[30] Erik Hollnagel, Robert L Wears, and Jeffrey Braithwaite. "From Safety-I to Safety-II: a white paper". In: *The resilient health care net: published simultaneously by the University of Southern Denmark, University of Florida, USA, and Macquarie University, Australia* (2015).

[31] William Hutiri, Osvaldo Papakyriakopoulos, and Alan Xiang. "Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, 2024, pp. 359–376. DOI: 10.1145/3630106.3658911. URL: https://doi.org/10.1145/3630106.3658911.

[32] AI Safety Institute. *Frontier AI: Capabilities and Risks Discussion Paper*. Accessed December 2024. 2023. URL: https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper.

[33] ISO/IEC/IEEE. *Systems and Software Engineering - System Life Cycle Processes*. ISO/IEC/IEEE 15288:2015. 2015.

[34] Emma Jones and Becky Ghani. *What is a Foundation Model?* Accessed: December 2024. 2023.

[35] Marten H. L. Kaas et al. "Ethics in conversation: Building an ethics assurance case for autonomous AI-enabled voice agents in healthcare". In: *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. TAS '23. Edinburgh, United Kingdom: Association for Computing Machinery, 2023. ISBN: 9798400707346. DOI: 10.1145/3597512.3599713. URL: https://doi.org/10.1145/3597512.3599713.

[36] Pickett Kate, Oliver James, and Richard Wilkinson. "Income Inequality and the Prevalence of Mental Illness: A Preliminary International Analysis". In: *Journal of Epidemiology and Community Health* 60.7 (2006), pp. 646–647. DOI: https://doi.org/10.1136/jech.2006.046631.

[37] John C. Knight. "Safety Critical Systems: Challenges and Directions". In: *Proceedings of the 24th International Conference on Software Engineering (ICSE '02)*. ACM, 2002, pp. 547–550. DOI: 10.1145/581339.581406. URL: https://doi.org/10.1145/581339.581406.

[38] Leonie Koessler and Jonas Schuett. *Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries*. 2023. arXiv: 2307.08823. URL: https://arxiv.org/abs/2307.08823.

[39] Daria Kuss and Mark Griffiths. "Social Networking Sites and Addiction: Ten Lessons Learned". In: *International Journal of Environmental Research and Public Health* 14.3 (2017). DOI: 10.3390/ijerph14030361.

[40] Seth Lazar and Alondra Nelson. "AI Safety on Whose Terms?" In: *Science* 381 (2023), p. 138. DOI: 10.1126/science.adi8982. URL: https://doi.org/10.1126/science.adi8982.

[41] Adam Leventhal et al. "Digital Media Use and Suicidal Behavior in U.S. Adolescents, 2009–2017". In: *Preventive Medicine Reports* 23 (2021).

[42] Nancy Leveson. *An Introduction to System Safety*. 2008. URL: https://appel.nasa.gov/2008/06/01/an-introduction-to-system-safety.

[43] Frank R. Manuele. "Acceptable Risk: Time for SH&E Professionals to Adopt the Concept". In: *Professional Safety* 55.6 (2010), pp. 30–38.

[44] Claudio Marino et al. "The Associations Between Problematic Facebook Use, Psychological Distress, and Well-Being Among Adolescents and Young Adults: A Systematic Review and Meta-Analysis". In: *Journal of Affective Disorders* 226 (2018), pp. 274–281. DOI: 10.1016/j.jad.2017.10.007.

[45] Sean McCarthy-Jones. "The Autonomous Mind: The Right to Freedom of Thought in the Twenty-First Century". In: *Frontiers in Artificial Intelligence* 2 (2019). DOI: 10.3389/frai.2019.00019. URL: https://doi.org/10.3389/frai.2019.00019.

[46] Brent Mittelstadt. "Principles Alone Cannot Guarantee Ethical AI". In: *Nature Machine Intelligence* 1 (2019), pp. 501–507. DOI: 10.1038/s42256-019-0114-4. URL: https://doi.org/10.1038/s42256-019-0114-4.

[47] Blake Montgomery. *Mother Says AI Chatbot Led Her Son to Kill Himself in Lawsuit Against Its Maker*. 2024.

[48] Blake Montgomery. *Why Silicon Valley Panicked Over Australia's Under-16 Social Media Ban*. Accessed: January 2025. 2024.

[49] Office for National Statistics. *Cost of Living and Depression in Adults, Great Britain: 29 September to 23 October 2022*. 2022. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/mentalhealth/articles/costoflivinganddepressioninadultsgreatbritain/29septemberto23october2022.

[50] Luisa Nazareno and Daniel Schiff. "The Impact of Automation and Artificial Intelligence on Worker Well-Being". In: *Technology in Society* 67 (2021), p. 101679. URL: https://doi.org/10.1016/j.techsoc.2021.101679.

[51] Jacqueline Nesi and Mitchell Prinstein. "Using Social Media for Social Comparison and Feedback-Seeking: Gender and Popularity Moderate Associations with Depressive Symptoms". In: *Journal of Abnormal Child Psychology* 43.8 (2015), pp. 1427–1438. DOI: 10.1007/s10802-015-0020-0.

[52] Ziad Obermeyer et al. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations". In: *Science* 366 (2019), p. 447. DOI: 10.1126/science.aax2342. URL: https://www.science.org/doi/full/10.1126/science.aax2342.

[53] OECD. *Measuring Population Mental Health*. 2023. URL: https://doi.org/10.1787/5171eef8-en.

[54] OpenAI. *Introducing ChatGPT*. 2025. URL: https://openai.com/index/chatgpt.

[55] World Health Organization. *Burn-out an 'Occupational Phenomenon': International Classification of Diseases*. 2019. URL: https://www.who.int/news/item/28-05-2019-burn-out-an-occupational-phenomenon-international-classification-of-diseases.

[56] University of Oxford. *Expert Comment: Oxford AI Experts Comment on the Outcomes of the UK AI Safety Summit*. Accessed: December, 2024. 2023. URL: https://www.ox.ac.uk/news/2023-11-03-expert-comment-oxford-ai-experts-comment-outcomes-uk-ai-safety-summit.

[57] Burak Ozturk et al. "Predicting Progression of Type 2 Diabetes Using Primary Care Data with Machine Learning". In: *Stud Health Technol Inform.* 302.38-42 (2023).

[58] UK Parliament. *Online Safety Act 2023 (c.50)*. 2023. URL: https://www.legislation.gov.uk/ukpga/2023/50.

[59] Iyad Raji and Raji Dobbe. "Concrete Problems in AI Safety, Revisited". In: (2023). eprint: arXivpreprint. URL: https://arxiv.org/abs/2401.10899.

[60] Rahul Rao. *The Dutch tax authority was felled by AI – What comes next?* IEEE Spectrum. 2022. URL: https://spectrum-ieee%20org.cdn.ampproject.org/c/s/spectrum.ieee.org/amp/artificial-intelligence-in-government-2657286505.

[61] Richard Ren et al. *Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?* 2024. arXiv: 2407.21792. URL: https://arxiv.org/abs/2407.21792.

[62] O. Renn and A. Klinke. "Systemic Risks: A New Challenge for Risk Management". In: *EMBO Reports* 5.Spec No(Suppl 1) (2004), S41–46. DOI: https://doi.org/10.1038/sj.embor.7400227.

[63] Royal Commission into the Robodebt Scheme. *Report*. 2023. URL: https://robodebt.royalcommission.gov.au/publications/report.

[64] Tahereh Saheb, Mouwafac Sidaoui, and Bill Schmarzo. "Convergence of Artificial Intelligence with Social Media: A Bibliometric & Qualitative Analysis". In: *Telematics and Informatics Reports* 14 (2024). DOI: https://doi.org/10.1016/j.teler.2024.100146.

[65] Conrad Sanderson et al. "Resolving Ethics Trade-offs in Implementing Responsible AI". In: *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 1208–1213. DOI: 10.1109/cai59869.2024.00215. URL: http://dx.doi.org/10.1109/CAI59869.2024.00215.

[66] Robert Schnitzer et al. "AI Hazard Management: A Framework for the Systematic Management of Root Causes for AI Risks". In: *Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications: 1st International Conference on Frontiers of AI, Ethics, and Multidisciplinary Applications (FAIEMA)*. 2023. DOI: 10.1007/978-981-99-9836-4_27. URL: https://doi.org/10.1007/978-981-99-9836-4_27.

[67] Renee Shelby et al. "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. Montréal, QC, Canada: Association for Computing Machinery, 2023, pp. 723–741. ISBN: 9798400702310. DOI: 10.1145/3600211.3604673. URL: https://doi.org/10.1145/3600211.3604673.

[68] Beard Simon, Rowe Thomas, and James Fox. "An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards". In: *Futures* 115 (2020). DOI: 10.1016/j.futures.2019.102469.

[69] Social Media and Children 2024 Legislation. *Social Media and Children 2024 Legislation*. Available at: https://www.ncsl.org/technology-and-communication/social-media-and-children-2024-legislation (Accessed: 15 January 2025). 2024.

[70] Rajagopalan Srinivasan et al. "Recent developments towards enhancing process safety: Inherent safety and cognitive engineering". In: *Computers & Chemical Engineering* 128 (2019), pp. 1–15. DOI: 10.1016/j.compchemeng.2019.05.034. URL: https://doi.org/10.1016/j.compchemeng.2019.05.034.

[71] International Organization for Standardization. *Information Technology - Artificial Intelligence - Guidance on Risk Management*. ISO/IEC 23894:2023. 2023. URL: https://www.iso.org/standard/77304.html.

[72] Risto Uuk et al. *A Taxonomy of Systemic Risks from General-Purpose AI*. 2024. eprint: https://arxiv.org/abs/2412.07780.

[73] Candice E Walker et al. "Effects of Social Media Use on Desire for Cosmetic Surgery Among Young Women". In: *Current Psychology* 40.7 (2021), pp. 3355–3364. URL: https://doi.org/10.1007/s12144-019-00282-1.

[74] Xiaoyang Wang et al. "Algorithmic Discrimination: Examining Its Types and Regulatory Measures with Emphasis on US Legal Practices". In: *Frontiers in Artificial Intelligence* 7.1320277 (2024). DOI: 10.3389/frai.2024.1320277. URL: https://doi.org/10.3389/frai.2024.1320277.

[75] Laura Weidinger et al. "Taxonomy of Risks posed by Language Models". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 214–229. ISBN: 9781450393522. DOI: 10.1145/3531146.3533088. URL: https://doi.org/10.1145/3531146.3533088.

[76] Christopher Wickens et al. *Engineering Psychology and Human Performance*. 5th. Routledge, 2018.

[77] Jing Xu et al. *Recipes for Safety in Open-domain Chatbots*. 2021. arXiv: 2010.07079. URL: https://arxiv.org/abs/2010.07079.