nature machine intelligence

Review article

https://doi.org/10.1038/s42256-025-01020-y

AI safety for everyone

Received: 28 August 2024

Accepted: 19 February 2025

Published online: 17 April 2025

Check for updates

Bálint Gyevnár[1,3] & Atoosa Kasirzadeh[2,3] ✉

Recent discussions and research in artificial intelligence (AI) safety have increasingly emphasized the deep connection between AI safety and existential risk from advanced AI systems, suggesting that work on AI safety necessarily entails serious consideration of potential existential threats. However, this framing has three potential drawbacks: it may exclude researchers and practitioners who are committed to AI safety but approach the field from different angles; it could lead the public to mistakenly view AI safety as focused solely on existential scenarios rather than addressing a wide spectrum of safety challenges; and it risks creating resistance to safety measures among those who disagree with predictions of existential AI risks. Here, through a systematic literature review of primarily peer-reviewed research, we find a vast array of concrete safety work that addresses immediate and practical concerns with current AI systems. This includes crucial areas such as adversarial robustness and interpretability, highlighting how AI safety research naturally extends existing technological and systems safety concerns and practices. Our findings suggest the need for an epistemically inclusive and pluralistic conception of AI safety that can accommodate the full range of safety considerations, motivations and perspectives that currently shape the field.

The rapid development and deployment of AI systems has made questions of safety increasingly urgent, demanding immediate attention from policymakers and governance bodies as they face critical decisions about regulatory frameworks, liability standards and safety certification requirements for AI systems. Recent narratives and research in AI safety have increasingly emphasized the deep connection between AI safety and existential risk from advanced AI systems[1–3]. This disparity has led to a divisive and sometimes unhealthy atmosphere, with some researchers even questioning the unique contribution of the 'AI safety' community[4]. This concentrated attention on existential risk has emerged despite—and perhaps overshadowed—decades of engineering and technical progress in building robust and reliable AI systems.

Historically, technological and systems safety research has evolved alongside each major industrial and computational advance. From ensuring aircraft structural integrity[5,6] to pharmaceutical security[7,8], and later expanding to cyber and internet safety[9,10], the field of technological and systems safety has consistently evolved in reaction to new technological paradigms. This evolution has produced robust engineering practices and governance frameworks[11–13] that could remain relevant to the challenges of AI development and deployment[13–15].

In this Review, we conduct a systematic review of primarily peer-reviewed research on AI safety to contextualize AI safety within broader technological and systems safety practice. Our analysis examines two key questions: (1) What categories of safety risks are addressed at different stages of an AI system life cycle, from development to deployment? (2) What concrete strategies and methodologies have researchers proposed to mitigate these identified risks? Our findings show that peer-reviewed research on AI safety has included a broad spectrum of concerns throughout AI development and deployment. The mathematical methods, computational tools, and algorithms developed for safe AI address fundamental challenges in current systems, from adversarial robustness to interpretability.

Previous attempts to bridge different safety concerns have often distinguished between concrete, near-term problems and broader, long-term existential challenges[16,17]. However, our literature review

[1]School of Informatics, University of Edinburgh, Edinburgh, UK. [2]Departments of Philosophy and Software and Societal Systems, Carnegie Mellon University, Pittsburgh, PA, USA. [3]These authors contributed equally: Bálint Gyevnár, Atoosa Kasirzadeh. ✉e-mail: atoosa.kasirzadeh@gmail.com

reveals that this dichotomy may oversimplify the rich landscape of AI safety research. The shared vocabulary and overlapping technical challenges—particularly evident in reinforcement learning (RL) paradigms where concepts such as corrigibility and adversarial robustness span multiple time horizons—indicate the value of a more integrated approach.

Our findings suggest that grounding AI safety discussions in the broader foundations of technological and system safety research could both enrich current debates and provide more robust guidance for policy and governance decisions. While existential and extreme risks from AI remain an important consideration, the field's historical roots in systems engineering and safety practices offer valuable views for addressing immediate challenges to build longer-term solutions. This suggests the need for an expanded discourse that reintegrates traditional safety engineering approaches with contemporary AI challenges, rather than treating existential risk as the primary lens through which to view AI safety. By reconnecting with these foundational conceptions, the field may be better equipped to address both immediate safety challenges and longer-term concerns, while maintaining the rigorous technical and theoretical standards that characterize effective safety research.

Limitations

Certain limitations must be considered when interpreting our results. First, AI safety research is a dynamically evolving field with continuous developments across numerous venues and fields. New research entities are emerging, conducting studies within and outside organizations, including research outputs from AI companies such as Anthropic, OpenAI and other relevant institutions. A one-off systematic literature review, such as ours, captures the state of the art only at the moment of querying, which, in our case, was on 1 November 2023. This means that our findings may not fully reflect the most recent advancements in AI safety research. Second, our focus on peer-reviewed research, while ensuring a standard of quality and rigour, inherently excludes an important portion of the AI safety landscape. Preprint repositories such as arXiv often serve as the first outlet for AI safety research, and their exclusion in a systematic review may result in overlooking important contributions. We attempted to mitigate this limitation through snowball sampling of high-impact arXiv papers, although this approach is not comprehensive and carries a risk of missing relevant publications. Third, our review process involved annotating each selected paper with relevant keywords beyond those provided by the authors. This approach may introduce some degree of annotator bias. However, given the substantial volume of papers selected (383 in total), we believe that the impact of this bias on our overall conclusions is minimal. Finally, our review does not include the extensive discussions found in online forums such as LessWrong or the AI Alignment Forum, which are popular platforms for discussing AI safety topics within certain communities. A comprehensive analysis of AI safety research would ideally include a detailed content analysis of these platforms and other non-peer-reviewed documents. Therefore, it is important to interpret our findings with caution, acknowledging that they provide a valuable but not exhaustive overview of the current state of AI safety research. Future research could expand on our work by incorporating a wider range of sources, including other preprint sources and non-peer-reviewed literature, to gain a more comprehensive understanding of the evolving landscape of AI safety. Despite these limitations, our findings provide a valuable snapshot of peer-reviewed AI safety research and motivate the need for a more inclusive understanding of the field's perceived scope and motivations. The novelty of our paper is leveraging empirical investigations in clarifying contested discussions about AI safety. Our review offers insights into research outputs, communities and potential avenues for fostering healthy research development. In the concluding section, we outline key next steps and open questions to guide future research.

Systematic review methodology

We conduct a systematic literature review of primarily peer-reviewed AI safety research. We have publicly released a dataset of the selected papers' metadata in various formats including our annotations and the code used to analyse and visualize these data after our review process under the following repository: https://github.com/gyevnarb/ai-safety-review.

Our review follows the guidelines outlined by Kitchenham and Charters[18], complemented by snowball sampling as recommended by Wohlin[19]. This approach combines a structured and repeatable methodology, as represented in peer-reviewed and indexed papers, with a targeted technique to capture emerging research not yet fully represented in the peer-reviewed literature.

We conducted our review and analysis to investigate two key research questions (RQs) in relation to the peer-reviewed published research on AI safety:

- RQ1: What types of risk related to the life cycle (design, development, deployment, operation and decommissioning) of AI systems are addressed in AI safety research?
- RQ2: What mitigation strategies—for example, concrete methods, design principles and governance recommendations—are proposed in recent AI safety research that directly address one or more of the above sources of AI risk?

Search and review process

To conduct our systematic review, we performed a multi-stage query process to identify relevant papers from the Web of Science and Scopus indexing databases. We chose these databases because they include not only computer science but also other peer-reviewed research that captures interdisciplinary work. We develop a hierarchy of increasingly refined queries for each research question, targeting the title, abstract and author keywords of papers. The querying process was finalized on 1 November 2023. We outline the query hierarchy below, using Web of Science notation:

- q1: AI $\vee$ AGI $\vee$ frontier AI $\vee$ artificial intelligence $\vee$ artificial (general $\vee$ super) intelligence $\vee$ (machine $\vee$ supervised $\vee$ unsupervised $\vee$ semi-supervised $\vee$ reinforcement) learning
- q2: safe* $\vee$ robust* $\vee$ align*
- q3: q1 $\wedge$ q2
- q4: q3 $\wedge$ (q4a $\vee$ q4b $\vee$ q4c $\vee$ q4d):
  - q4a: design $\vee$ develop* $\vee$ architecture $\vee$ model* $\vee$ framework $\vee$ system
  - q4b: deploy* $\vee$ distribut* $\vee$ data* $\vee$ train* $\vee$ fine-tun*
  - q4c: operat* $\vee$ interact* $\vee$ online
  - q4d: decomission* $\vee$ remov* $\vee$ eras* $\vee$ delet*

We start with a broad query (q1) that includes various terms related to AI and machine learning (ML), such as 'AI', 'AGI', 'frontier AI', 'artificial intelligence', 'machine learning' and so on (see q1 above for full details). This initial query aims to capture all papers potentially related to AI research. Next, we introduce a second query (q2) that incorporates high-level safety-related keywords of 'safe*', 'robust*' and 'align*'. The asterisks act as wildcards to capture variations of these terms (for example, 'safety', 'safer', 'safest', 'robustness' and so on). We then combine q1 and q2 to create q3, which selects papers relevant to AI or ML research while ensuring they also address safety-related concepts. This step narrows down the results to publications specifically focusing on AI safety. To further refine the selection and align it with the specific focus areas of our RQ1, we introduce q4. This query includes four subqueries, each corresponding to terms related to the four areas of RQ1: design, deployment, operation and decommissioning of AI systems. To ensure that we do not miss important non-peer-reviewed contributions, we supplement our systematic search with snowball sampling, starting with 12 seed papers identified as highly influential in the AI safety field[16,20–30]. This allows us to identify additional relevant papers that are not captured by the initial queries.

After removing duplicates, our query process yielded 2,666 papers from the database search and 117 papers from snowball sampling. We then conducted a two-stage review process, first filtering based on titles and abstracts, followed by a comprehensive full-text review. This resulted in a final set of 383 papers for our analysis. We applied the following exclusion criteria during both the title and abstract screening and the full-text review stages to determine a paper's eligibility for inclusion in our analysis: (1) focus: the paper does not primarily focus on AI or a directly related subfield; (2) motivation: the paper's stated motivation does not include a clear need for developing safe AI algorithms; (3): generalizability: the paper's contributions are specific to a very particular application domain (for example, flood prediction) and do not offer broader insights or methodologies applicable to AI safety in general.

Annotation process

As part of our review process, we recorded important metadata about each paper, which included the publication year, author affiliations and Google Scholar citation count (as of 27 January 2024). We then performed a thorough inductive coding process[31] across the selected papers. We enriched each paper's author-written keywords by adding relevant terms (that is, codes) based on a thorough reading of the full text. This expanded set of keywords encompassed both problem- and method-specific terms, as well as broader categorizations such as 'algorithm' (for papers that propose an algorithm), 'theoretical' (for purely theoretical contributions) and 'framework'. We also gradually refined and consolidated our set of keywords as we progressed through the papers. After gaining a holistic understanding of the selected publications, we further categorized each paper based on its assigned keywords according to its methodological approach, the specific safety risks it addressed, the types of risk mitigated by its proposed methods and the overarching category of its methodology.

Empirical findings

To begin, we present a high-level bibliometric overview of the trends and concepts prevalent in the AI safety literature we reviewed. Looking at the total number of publications per year in Fig. 1, we observe increasing growth since 2016, which we assume is partially caused by the extensive development and deployment of deep learning models. Nevertheless, this observation reinforces the need to look more deeply into the state of the field. We first look at a word cloud of salient terms to understand the most important concepts among our selected papers. We then analyse patterns that emerge from the co-occurrences of different terms in abstracts and titles.

Word cloud of salient terms

Figure 2 illustrates a word cloud of the most salient terms found in the abstracts of the selected papers, after morphological standardization. The terms are ranked by their tf-idf (term frequency-inverse document frequency) score, a metric that emphasizes both the frequency and distinctiveness of a term within the corpus. This allows us to highlight important, but potentially less frequent, terms while de-emphasizing common or redundant words. The word cloud highlights several prominent themes within the AI safety literature. A notable portion of papers focus on safe RL, evidenced by terms such as 'robust', 'control', 'agent' and 'explore'. In addition, there is a strong emphasis on adversarial attacks, as indicated by the prevalence of terms such as 'adversarial' and 'attack'. Finally, domain adaptation emerges as another significant area of concern, with terms such as 'domain', 'distribution' and 'adapt' appearing frequently.

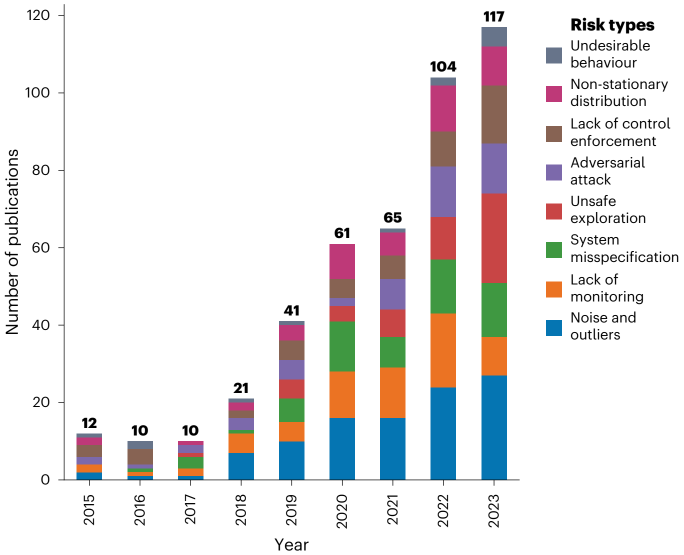

Fig. 1 | Number of publications over time since 2015 for the different risk types that were identified in the paper as a result of a systematic literature review. Some papers may belong to more than one risk type.

Term co-occurrence graph

To gain deeper qualitative insights into the landscape of peer-reviewed AI safety research, we constructed a co-occurrence graph of terms appearing in the abstracts and titles of selected papers using the VOSViewer tool[32]. Figure 3 depicts this graph. We observe four distinct clusters.

The motivations and concerns driving AI safety research is shown in the largest, red cluster. We observe a clear emphasis on both the development and deployment of AI systems, especially in relation to humans, health and society. Here the focus is human- and society-centric aspects of AI safety, including trust, accountability and safety assurance.

Safe RL and related terms are aggregated in the blue cluster. Research in this cluster primarily centres on the safe control of agents under constraints in uncertain dynamic environments. While most studies test methods in simulated environments, there is also considerable work on the safety of nonlinear systems employing RL for optimal control. Overall, the focus tends towards mathematically well-defined problems such as optimality, convergence and stability, sample efficiency, and constraint satisfaction.

Supervised learning, primarily classification tasks, is the predominant focus of the green cluster. This cluster features research on methods such as neural networks, extreme learning machines, support vector machines and random forests. Studies prioritize addressing issues such as robustness to noise and outliers, generalization performance, and accuracy.

Finally, adversarial attacks and defence is highlighted in the yellow cluster, especially as they relate to deep neural networks. The existence of a separate cluster for this class of problems emphasizes the importance of this method for AI safety. Methods here focus on robust-to-outlier optimization, adversarial training through synthetic datasets and learning robust representations via semi-supervised learning.

The four clusters identified in the co-occurrence graph offer a strong reason for viewing AI safety research as a natural extension of traditional technological safety practices. The centrality of human and societal well-being in the red cluster aligns with the core purpose of technological safety: safeguarding human life and minimizing harm. This focus echoes traditional safety concerns in fields such as aviation[6,33] or medicine[34,35], with an emphasis on protecting users and the public. The emphasis on mathematically rigorous control of

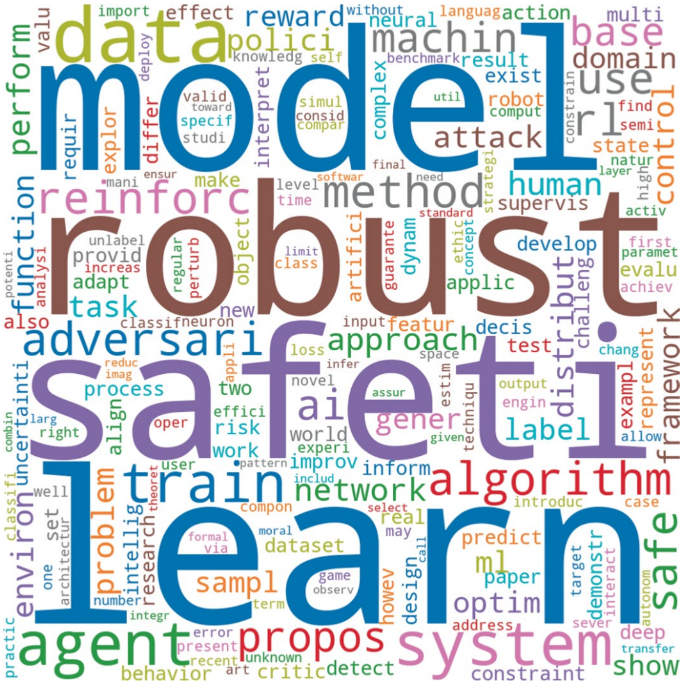

Fig. 2 | World cloud of morphologically standardized terms occurring in the abstracts and titles of selected papers weighted by their tf-idf score. We excluded stop words from the analysis.

agents under constraints in uncertain environments in the second cluster directly mirrors the principles of control theory and systems engineering, which are foundational to ensuring the safety of complex systems[11,36]. The focus on optimality, stability and constraint satisfaction reflects the fundamental goals of traditional safety engineering, such as preventing failures and ensuring predictable behaviour. In the green cluster, the emphasis on robustness, generalization and accuracy in supervised learning reflects the core concerns of reliability engineering. Just as engineers strive to build robust and reliable physical systems, AI safety researchers are focused on creating ML models that can perform consistently and accurately in diverse real-world scenarios. The yellow cluster shows the growing recognition of cybersecurity as a critical component of technological safety[37,38]. The focus on adversarial attacks and defence strategies mirrors the evolving challenges of protecting digital infrastructure and data integrity.

Types of safety risk

To answer RQ1 and better understand the motivations behind research on AI safety, it is important to examine the various types of risk addressed in research on AI safety. Drawing upon Google DeepMind's categorization of AI safety risks[39], our review identifies eight overarching risk types, as summarized in Fig. 4 and Table 1.

The safety risks in increasing order of frequency are: undesirable behaviour, non-stationary distributions, adversarial attacks, unsafe exploration, lack of control enforcement, system misspecification, lack of monitoring, and noise and outliers.

The largest group of papers focuses on risk stemming from noise and outliers. Recurring challenges in this area include brittle representations[40], low classification performance[41], a lack of generalization in the presence of noisy labels[42], outliers[43], or input perturbations and corruptions[44].

The current literature also studies the significant lack of monitoring of AI systems. Research addressing this risk ranges from theoretical violations of ethical and safety principles[45] to privacy violations[46]. The 'black box' nature of popular deep learning systems has also drawn considerable attention owing to its potential to diminish human agency and obstruct understanding of the system's internal workings and memory. Consequently, there has been a surge of interest in reverse-engineering ML models[47], developing interpretable representation learning[48] and advancing explainable AI[49,50].

System misspecification, or misunderstanding the requirements and purpose of AI-based systems, has also garnered significant attention. Potential risks include incorrectly eliciting requirements for AI systems' capabilities[51], making suboptimal modelling choices[52] or choosing inappropriate hyperparameters[53]. In addition, methods may fail to adapt to the changing requirements of their domain[54], particularly due to the slow pace of retraining AI models[55]. Privacy violations due to inadequate ethical data management also fall under this risk category[56].

Attention has also been given to risks that originate from a lack of control enforcement. A primary research focus here is the misalignment of agent goals with human preferences[25], which can lead to wireheading[57], mesa-optimization (that is, an optimizer creating another optimizer)[58] and other undesirable emergent behaviours[59]. There is also a focus on the lack of systemic safety, where the goal is to ensure that the AI system is safe in the broader context of its deployment[23,60].

A notable portion of research focuses on the unsafe exploration of autonomous agents, primarily in the context of RL[61], constraint violations[62], unintended behaviour from incorrect domain or reward specification[53], and incorrect behaviour due to continuous deployment and learning[63,64].

Adversarial attacks constitute another significant risk, addressed by methods that aim to create or detect such attacks[65] or leverage adversarial samples for more robust training[66]. The risk of poisoned training data is also a concern[67,68]. Adversarial attacks not only compromise the robustness and generalization performance of AI systems but also can introduce a backdoor[69], which can potentially be exploited by malicious actors.

The challenges of non-stationary distributions, where distributions change over time, have also been explored. These studies address issues such as behaviour in the presence of out-of-distribution samples[70], non-stationary environments in RL[71], partial information[72] and domain adaptation[54]. The overarching goal is to ensure safe behaviour and limit the consequences of unsafe actions, even in novel or unforeseen situations.

The final group of risk types is concerned with undesirable behaviour of AI systems. Research has looked into mathematically proving certain unfavourable outcomes of utility-maximizing rational agents, where issues include self-modifying[73] and wireheading agents that by-pass reward signals to maximize their own utility[57], and corrigibility of uncooperative agents[74]. In addition, there are some papers that directly discuss the existential risk of AI systems as a fundamental problem of AI safety. These papers present theoretical arguments regarding the emergence, design, and containment of malevolent and superintelligent AI[75–77].

These identified risk types in AI safety research closely mirror those found in established fields of technological safety research. In AI safety, the focus on ensuring the reliability and predictability of AI systems, including robustness to noise and outliers, generalization performance, and adaptability, mirrors concerns in engineering fields where reliability is of core concern, such as aviation[6] or nuclear power[78].

Furthermore, the research on adversarial attacks, control enforcement and safe exploration in AI safety directly translates to broader concepts of system robustness and control. The goal of designing systems that can withstand unexpected perturbations, resist malicious attacks and operate safely within defined boundaries is shared by both AI safety and traditional safety engineering. For example, the use of formal verification methods to ensure the safety of AI systems has roots in the verification of software and hardware systems[79].

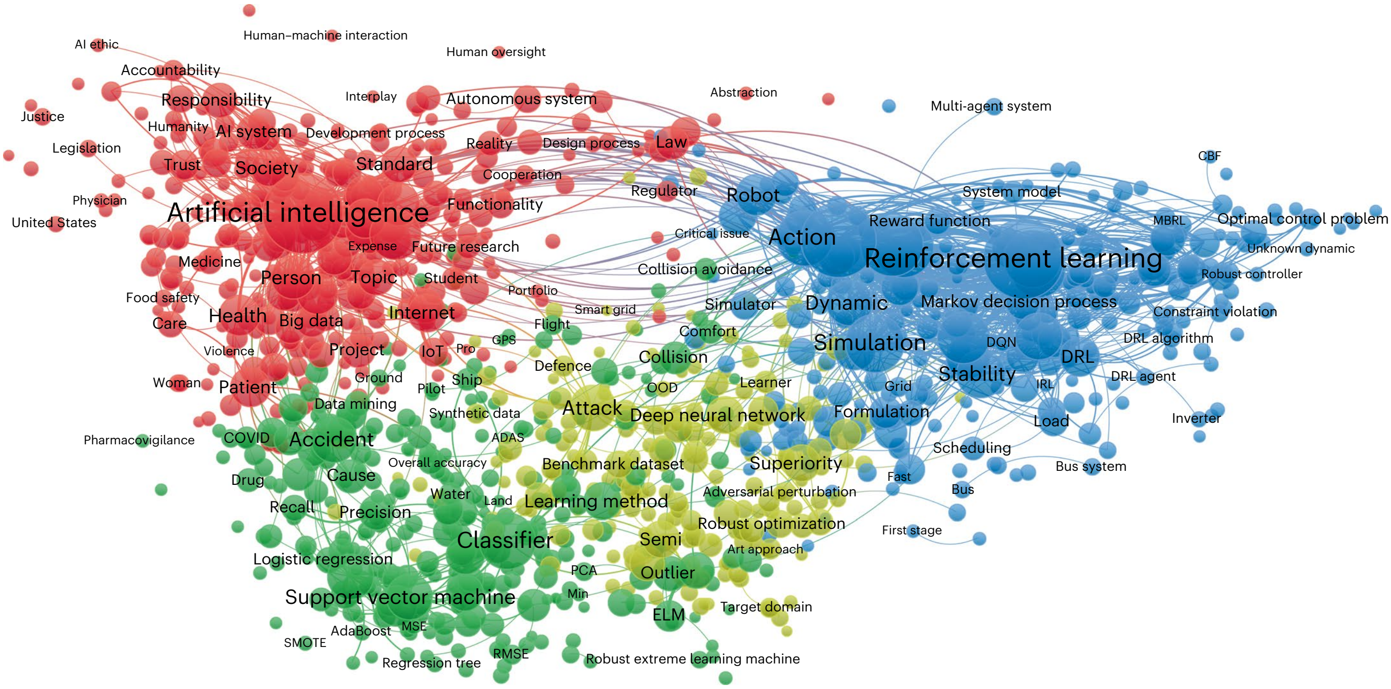

Fig. 3 | Graph of term co-occurrence in abstracts with binary counting, a minimum term frequency of 10 and a relevance score of at least 60%. The figure was produced using the VOSViewer tool[32]. Nodes represent terms and edges connect terms that co-occur within the same abstract or title. The size of each node reflects its relevance score as calculated by VOSViewer. Owing to the large number of nodes, not all labels are shown in the figure. The red cluster denotes the motivations and concerns that drive AI safety research; the blue cluster represents safe RL and related terms; the yellow cluster covers adversarial attacks and defence; and the green cluster denotes supervised learning and classification tasks. AdaBoost, adaptive boosting; ADAS, advanced driver assistance system; CBF, control barrier function; DRL, deep reinforcement learning; DQN, deep Q-network; ELM, extreme learning machine; GPS, global positioning system; IoT, Internet of Things; IRL, inverse reinforcement learning; MBRL, model-based reinforcement learning; Min, minimum; MSE, mean squared error; OOD, out-of-distribution; PCA, principal component analysis; RMSE, root mean squared error; SMOTE, synthetic minority over-sampling technique.

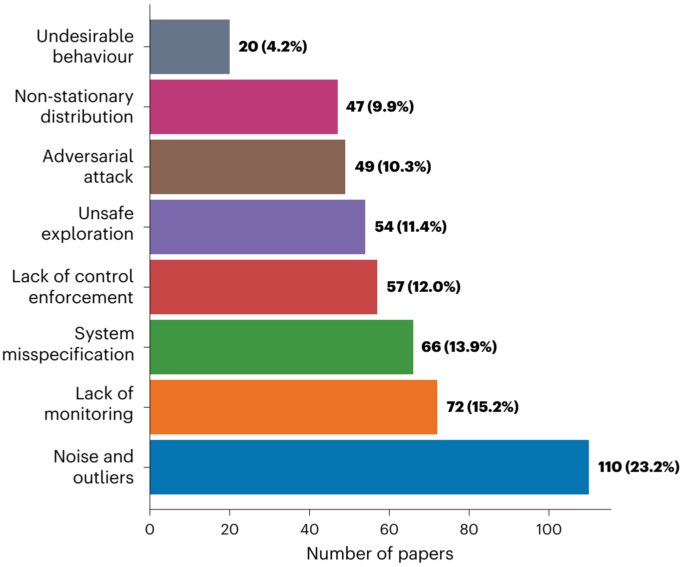

Fig. 4 | Number and percentages of the selected papers that address various risk types. Some papers may belong to multiple risk types so the number of papers add up to more than the total number of collected papers (383).

Developed AI safety methodologies

To address RQ2, we examine the concrete mitigation strategies proposed in recent AI safety research to directly address the aforementioned sources of AI risk: What mitigation strategies—for example, concrete methods, design principles and governance recommendations—are proposed in recent AI safety research that directly address one or more of the above sources of AI risk?

We see a dynamic landscape of AI safety research, including both theoretical and applied approaches (Fig. 5). We consider a work as theoretical if its primary conclusions stem from theoretical analysis, mathematical proofs, philosophical arguments, literature reviews or conceptual frameworks that are not empirically validated. Conversely, a work is classified as applied if its conclusions are derived from empirical evidence and supported by data.

Our analysis shows the following trend: theoretical works in AI safety predominantly offer general frameworks and recommendations, for example, refs. 80–85, whereas applied works primarily focus on developing and testing concrete algorithms, for example, refs. 65,86–90. This dichotomy highlights a key distinction: theoretical research often prioritizes conceptual and methodological foundations, whereas applied research emphasizes practical implementation and testing. However, the applied algorithms could lack the rigorous guarantees that accompany analytical results found in theoretical work.

Beyond the theoretical/applied divide, we investigate the specific methodologies employed in AI safety research. Figure 6 illustrates the broad categories of methods proposed in the selected papers, each accompanied by representative citations based on citation counts to highlight their prevalence in the literature.

Applied algorithms research focuses on developing and empirically evaluating novel algorithms to enhance the safety of ML systems, such as supervised and unsupervised learning. These constitute the largest group within AI safety research and cover a wide range of techniques, including: robust classification algorithms resilient to noise[41,91,92], adversarial attacks[66,89,93], machine unlearning techniques for deep neural networks[56,94,95], and methods for improving domain generalization[96–98].

Table 1 | Representative contributions with concrete methods addressing problems for each risk type

Risk	Ref.	Contribution
Undesirable behaviour	73,74	Corrigibility of rational, utility-based agents What happens when an agent can modify its own code?
Non-stationary distributions	172,173	Domain-adversarial training of robust neural networks Domain generalization via meta-regularization
Adversarial attack	127,174	Qualitative analysis of adversarial perturbations Adversarial robustness as robust optimization
Unsafe exploration	118,175	Benchmarking safe exploration in deep RL Safe exploration for interactive ML
Lack of control enforcement	86,176	Large language models with human feedback to follow instructions Imitation learning via inverse RL
System misspecification	55,56	Efficient machine unlearning Rapid retraining of ML models
Lack of monitoring	114,128	Sanity checks for interpreting saliency maps Concept activation vectors in convolutional neural networks for interpretability
Noise and outliers	123,177	Baseline for detecting misclassified/ out-of-distribution samples Benchmarking robustness to corruptions/perturbations

Publications were picked by citation count.

Agent simulations, complementary to applied algorithms, focus on designing and evaluating safer agent training algorithms, predominantly drawing on existing RL literature. While these simulations could be subsumed under the category of applied algorithms, we distinguish between the two to highlight the significant attention that safer agent learning receives. Research in this area includes various approaches such as constrained Markov decision processes[62,99,100], enforcement of hard constraints[101,102], model-based RL for safe exploration[44,103,104], reward learning and inverse RL[21,86,105], multi-agent RL with a focus on safety and cooperation[106–108], and RL with human oversight or feedback mechanisms for enhanced safety[20,109,110].

Notably, earlier research in agent simulations was primarily motivated by technical and computational challenges of RL. However, recent studies have expanded their focus to explicitly address value alignment, aiming to align the reward functions of machines with human goals. Interestingly, despite the potential for real-world impact, research on real-world testing of embodied AI systems remains relatively limited. These studies address crucial safety concerns that may be overlooked in the design of unembodied AI systems, ranging from contact-safe continuous control to collaborative robotics[111–113].

Analysis frameworks is the third most prevalent category. These works are predominantly concerned with offering frameworks for exploring and evaluating the vulnerabilities of existing AI systems[114–117], proposing benchmark tasks and environments for assessing AI safety[118–120], and safety verification processes[60,121,122]. We include datasets in this category[23,123,124] as they provide standardized frameworks for evaluating AI systems, even though they are not frameworks in the strictest sense.

Mechanistic interpretability methods are a related category that investigates the inner workings of deep learning models to uncover the causal mechanisms behind their decision-making processes[47,125,126]. This often aims to develop automated interpretability tools for auditing and ensuring safety. Although the term 'mechanistic interpretability' has gained recent prominence, various forms of interpretability methods have been present since the rise of deep learning, often under the umbrella of explainable AI techniques[127–131].

Approximately 10% of the selected papers presented theoretical algorithms with analytical proofs, often foregoing empirical evaluation. These studies typically approached AI safety from the perspective of verification, using variance or error bounds[53,132,133], constraint satisfiability[102], or by demonstrating desirable properties such as metric goodness or Lyapunov functions within the problem set-up[134,135]. A notable portion of the papers also focused on proposing various design frameworks. These included methodologies for eliciting safety and system requirements[52,136,137], hierarchical integration of ML systems[138–140] and actionable ethical design processes[141–143].

A substantial portion of AI safety research consists of literature reviews, examining both well-established areas[16,144,145] and emerging or hypothetical issues[30,146,147].

Theoretical frameworks provide high-level analyses of AI safety issues, often by characterizing desirable properties of human–AI interactions[85,148,149], advocating for alternative approaches[150,151] or formalizing existing but vaguely defined concepts[84,152,153]. Within this category, a significant body of work employs mathematical reasoning to explore the safety of artificial general intelligence in the context of rational agents[73,74,81,154].

The final category includes purely philosophical research, exploring diverse questions and perspectives often related to value alignment[82,155,156] and the theorized risks of artificial general intelligence[59,157,158], responsible AI deployment[45,80,159] and the legal personhood of AIs[160].

The distribution of methodologies in AI safety research positions AI safety as an organic evolution of systems and technological safety practices. The abundance of research on developing and empirically evaluating algorithms for robust classification, adversarial defence and machine unlearning directly reflects the core of engineering practice: creating practical solutions to real-world problems. This emphasis on empirical validation and the focus on improving existing ML techniques resonate with the iterative and improvement-oriented nature of traditional safety engineering. Research on safer agent training algorithms, drawing heavily from the RL literature, aligns with the established practice of using simulations to test and refine safety-critical systems in controlled environments[161]. This approach allows researchers to explore potential risks and develop mitigation strategies before deploying AI systems in real-world scenarios, similar to flight simulators used in aviation safety. Although limited compared with other categories,

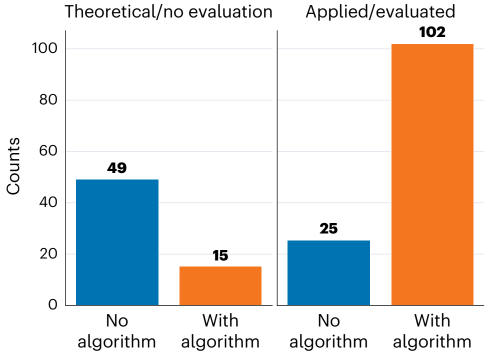

Fig. 5 | Comparison of papers that provide only theoretical results without significant empirical testing against those that give sufficient evaluation of their proposed system. Papers are further grouped by whether they propose a concrete algorithm. Note that the number of papers overlap as some propose solutions in more than one category.

https://doi.org/10.1038/s42256-025-01020-y

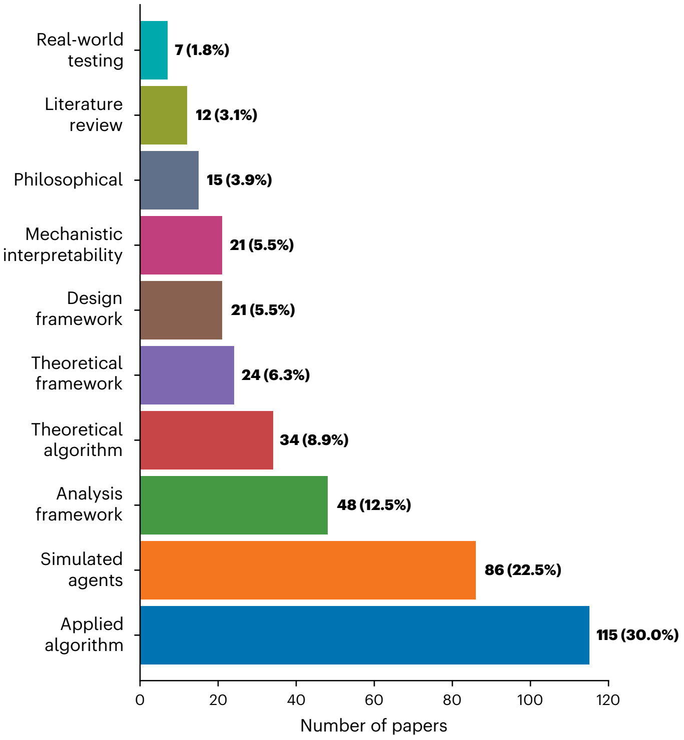

Fig. 6 | Distribution of categories of research undertaken to categorize and address the sources of risks. Some papers may belong into multiple categories.

the research on real-world testing of embodied AI systems shows the importance of validating theoretical models in practical settings.

This emphasis on real-world application is consistent with systems safety engineering practices that prioritize testing and validation in operational environments to ensure the safety and reliability of complex systems. The prevalence of research on analysis frameworks, datasets and mechanistic interpretability highlights the growing emphasis on transparency, explainability and accountability in AI safety. The focus on understanding the inner workings of AI systems and developing tools for evaluating their behaviour mirrors the rigorous analysis and testing protocols used in safety engineering to identify and mitigate safety risks.

Discussion and future research

In this paper, we conducted a systematic review of primarily peer-reviewed literature to unpack the diverse technical and practical challenges encompassed in AI safety. Our empirical analysis shows a broad landscape of motivations and outcomes driving AI safety research. The significance of these motivations and research outcomes comes from a desire to ensure that the AI systems we are building are reliable, trustworthy and beneficial for society. By examining a diverse body of peer-reviewed literature, we have found that AI safety research addresses a wide range of risks across most of the life cycle of AI systems. These risks echo a variety of concerns, including design misspecification, lack of robustness, inadequate monitoring and potential biases embedded within AI systems. Design misspecification can lead to the AI systems that behave in unintended and potentially harmful ways, while a lack of robustness can make them vulnerable to errors and malfunctions. Inadequate monitoring can prevent the detection and correction of issues.

The framing of AI safety primarily through existential risk[1–3] creates three problems: (1) it sidelines researchers who approach safety from other critical angles; (2) it misleads the public into thinking AI safety matters only for extreme scenarios rather than current systems; and (3) it generates resistance to AI safety practices from those who disagree with existential risk predictions. Just as other fields of engineering and technology have developed robust safety practices to mitigate risks and ensure the safe operation of complex systems, so too can AI safety research be viewed as an integral part of the progression within the broader domain of systems and technological safety. Approaching AI safety as an instance of systems safety practice has important implications for the field of AI safety[13–15]. First, by recognizing the relevance of AI safety to a diversity of risks, we can expand the circle of stakeholders involved in the discourse about AI safety. This expanded engagement can lead to increased funding and support for AI safety research, as well as a shift in the discourse towards more practical and inclusive solutions. Second, we think that a more epistemically inclusive research environment helps discussions to demystify existential risks from fresh perspectives[1].

While our argument advocates for a broader perspective on AI safety, it is crucial to emphasize that we do not intend to diminish the importance of critically engaging with existential risk concerns. These risks remain a significant area of research[1,162–165] as well as public and policy debate[166–169], and our findings do not negate the need for continued investigation and dialogue within this domain. Rather than diminishing these concerns, we think that incorporating a scientifically grounded, broader range of AI safety considerations will strengthen our collective ability to develop and deploy AI systems responsibly.

Several research questions remain open. We highlight two of them below.

First, the concept of 'sociotechnical AI safety'[2] merits deeper examination. While recent work has advanced our understanding of this approach in AI evaluation[170] and language model alignment[171], we still need a more comprehensive understanding of how sociotechnical frameworks can enhance AI safety efforts.

Second, while our review focused on peer-reviewed literature, future research should expand its scope to include non-peer-reviewed sources, such as more preprint papers, technical reports, and substantive research content on online forums and workshops. This broader analysis would provide a more comprehensive understanding of the diverse perspectives and approaches within AI safety and identify areas for future research and collaboration. Our empirical analysis is just a first step in understanding a complex practice.

References

1. Kasirzadeh, A. Two types of AI existential risk: decisive and accumulative. *Philos. Stud.* https://doi.org/10.1007/s11098-025-02301-3 (2025).
2. Lazar, S. & Nelson, A. AI safety on whose terms? *Science* **381**, 138–138 (2023).
3. Ahmed, S., Jaźwińska, K., Ahlawat, A., Winecoff, A. & Wang, M. Field-building and the epistemic culture of AI safety. *First Monday* https://doi.org/10.5210/fm.v29i4.13626 (2024).
4. Bender, E. M. Talking about a 'schism' is ahistorical. *Medium* https://medium.com/@emilymenonbender/talking-about-a-schism-is-ahistorical-3c454a77220f (2023).
5. Krause, S. S. *Aircraft Safety* (McGraw-Hill, 2003).
6. Boyd, D. D. A review of general aviation safety (1984–2017). *Aerosp. Med. Hum. Perform.* **88**, 657–664 (2017).
7. Pifferi, G. & Restani, P. The safety of pharmaceutical excipients. *Farmaco* **58**, 541–550 (2003).
8. Leveson, N. et al. Applying system engineering to pharmaceutical safety. *J. Healthc. Eng.* **3**, 391–414 (2012).
9. De Kimpe, L., Walrave, M., Ponnet, K. & Van Ouytsel, J. Internet safety. In *The International Encyclopedia of Media Literacy* (eds Hobbs, R. & Mihailidis, P.) (Wiley, 2019).
10. Salim, H. M. *Cyber Safety: A Systems Thinking and Systems Theory Approach to Managing Cyber Security Risks*. PhD thesis, Massachusetts Institute of Technology (2014).

11. Leveson, N. G. *Engineering a Safer World: Systems Thinking Applied to Safety* (MIT Press, 2016).
12. Varshney, K. R. Engineering safety in machine learning. In *Information Theory and Applications Workshop (ITA)* 1–5 (IEEE, 2016).
13. Rismani, S. et al. From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ML. In *Proc. 2023 CHI Conference on Human Factors in Computing Systems* 1–18 (2023).
14. Dobbe, R. System safety and artificial intelligence. In *FAccT '22: Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 1584–1584 (ACM, 2022).
15. Rismani, S. et al. Beyond the ML model: applying safety engineering frameworks to text-to-image development. In *AIES '23: Proc. 2023 AAAI/ACM Conference on AI, Ethics, and Society* 70–83 (ACM,2023).
16. Amodei, D. et al. Concrete problems in AI safety. Preprint at https://doi.org/10.48550/arXiv.1606.06565 (2016).
17. Raji, I. D. & Dobbe, R. Concrete problems in AI safety, revisited. Preprint at https://doi.org/10.48550/arXiv.2401.10899 (2023).
18. Kitchenham, B. & Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering* EBSE Technical Report EBSE-2007-01 (School of Computer Science and Mathematics, Keele University, 2007).
19. Wohlin, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering In *EASE '14: Proc. 18th International Conference on Evaluation and Assessment in Software Engineering* Article 38, 1–10 (ACM, 2014).
20. Irving, G., Christiano, P. & Amodei, D. AI safety via debate. Preprint at https://doi.org/10.48550/arXiv.1805.00899 (2018).
21. Ng, A. Y. & Russell, S. J. Algorithms for inverse reinforcement learning. In *ICML'00: Proc. 17th International Conference on Machine Learning* 663–670 (ACM, 2000).
22. Elhage, N. et al. Toy models of superposition. Preprint at https://doi.org/10.48550/arXiv.2209.10652 (2022).
23. Hendrycks, D. et al. Aligning AI with shared human values. In *International Conference on Learning Representations* (ICLR, 2021).
24. Yampolskiy, R. V. Artificial intelligence safety and cybersecurity: a timeline of AI failures. Preprint at https://doi.org/10.48550/arXiv.1610.07997 (2016).
25. Hadfield-Menell, D., Russell, S. J., Abbeel, P. & Dragan, A. Cooperative inverse reinforcement learning. In *NIPS'16: Proc. 30th International Conference on Neural Information Processing Systems* 3916–3924 (ACM, 2016).
26. Xu, H., Zhu, T., Zhang, L., Zhou, W. & Yu, P. S. Machine unlearning: a survey. *ACM Comput. Surv.* **56**, 9.1–9.36 (2023).
27. Russell, S., Dewey, D. & Tegmark, M. Research priorities for robust and beneficial artificial intelligence. *AI Mag.* **36**, 105–114 (2015).
28. Willers, O. et al. Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception tasks. In *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops: Lecture Notes in Computer Science* (eds Casimiro, A. et al.) 336–350 (Springer, 2020).
29. Mohseni, S. et al. Taxonomy of machine learning safety: a survey and primer. *ACM Comput. Surv.* **55**, 1–38 (2022).
30. Hendrycks, D., Carlini, N., Schulman, J. & Steinhardt, J. Unsolved problems in ML safety. Preprint at https://doi.org/10.48550/arXiv.2109.13916 (2022).
31. Boyatzis, R. E. *Transforming Qualitative Information: Thematic Analysis and Code Development* (Sage, 1998).
32. van Eck, N. J. & Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **84**, 523–538 (2010).
33. Oster, C. V. Jr, Strong, J. S. & Zorn, C. K. Analyzing aviation safety: problems, challenges, opportunities. *Res. Transport. Econ.* **43**, 148–164 (2013).
34. Donaldson, M. S., Corrigan, J. M. & Kohn, L. T. (eds). *To Err is Human: Building a Safer Health System* (National Academies Press, 2000).
35. Bates, D. W. et al. The safety of inpatient health care. *N. Engl. J. Med.* **388**, 142–153 (2023).
36. Marais, K., et al. *Beyond Normal Accidents and High Reliability Organizations: The Need for an Alternative Approach to Safety in Complex Systems* (Citeseer, 2004).
37. Griffor, E. *Handbook of System Safety and Security: Cyber Risk and Risk Management, Cyber Security, Threat Analysis, Functional Safety, Software Systems, and Cyber Physical Systems* (Syngress, 2016).
38. Prasad, R. & Rohokale, V. *Cyber Security: the Lifeline of Information and Communication Technology* (Springer, 2020).
39. Ortega, P. A., Maini, V. & the DeepMind Safety Team. Building safe artificial intelligence: specification, robustness, and assurance. *Medium* https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1 (2018).
40. Meng, Y. et al. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 10367–10378 (ACL, 2021).
41. Wang, K. & Guo, P. A robust automated machine learning system with pseudoinverse learning. *Cognit. Comput.* **13**, 724–735 (2021).
42. Cappozzo, A., Greselin, F. & Murphy, T. B. A robust approach to model-based classification based on trimming and constraints: semi-supervised learning in presence of outliers and label noise. *Adv. Data Anal. Classif.* **14**, 327–354 (2020).
43. Li, W. & Wang, Y. A robust supervised subspace learning approach for output- relevant prediction and detection against outliers. *J. Process Control* **106**, 184–194 (2021).
44. Curi, S., Bogunovic, I. & Krause, A. Combining pessimism with optimism for robust and efficient model-based deep reinforcement learning. In *Proc. 38th International Conference on Machine Learning* Vo. 139, 2254–2264 (PMLR, 2021).
45. Dobbe, R., Gilbert, T. K. & Mintz, Y. Hard choices in artificial intelligence. *Artif. Intell.* **300**, 103555 (2021).
46. Dwork, C. & Feldman, V. Privacy-preserving prediction. In *Proc. 31st Conference On Learning Theory* Vol. 75, 1693–1702 (PMLR, 2018).
47. Elhage, N. et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread* https://transformer-circuits.pub/2021/framework/index.html (2021).
48. Kim, H. & Mnih, A. Disentangling by factorising. In *Proc. 35th International Conference on Machine Learning* Vol. 80, 2649–2658 (PMLR, 2018).
49. Ward, F. & Habli, I. An assurance case pattern for the interpretability of machine learning in safety-critical systems. In *Computer Safety, Reliability, and Security. SAFECOMP 2020. Lecture Notes in Computer Science* (ed Casimiro, A. et al.) Vol. 12235, 395–407 (2020).
50. Gyevnar, B., Ferguson, N. & Schafer, B. Bridging the transparency gap: what can explainable AI learn from the AI Act? In *Frontiers in Artificial Intelligence and Applications* Vol. 372, 964–971 (IOS, 2023).
51. Reimann, L. & Kniesel-Wünsche, G. Safe-DS: a domain specific language to make data science safe. In *ICSE-NIER '23: Proc. 45th International Conference on Software Engineering: New Ideas and Emerging Results* 72–77 (ACM, 2023).
52. Dey, S. & Lee, S.-W. A multi-layered collaborative framework for evidence-driven data requirements engineering for machine learning-based safety-critical systems. In *SAC '23: Proc. 38th ACM/SIGAPP Symposium on Applied Computing* 1404–1413 (ACM, 2023).

53. Wei, C.-Y., Dann, C. & Zimmert, J. A model selection approach for corruption robust reinforcement learning. *Proc. Machine Learning Research* **167**, 1043–1096 (2022).
54. Ghosh, A., Tschiatschek, S., Mahdavi, H. & Singla, A. Towards deployment of robust cooperative AI agents: an algorithmic framework for learning adaptive policies. In *AAMAS '20: Proc. 19th International Conference on Autonomous Agents and MultiAgent Systems* 447–455 (ACM, 2020).
55. Wu, Y., Dobriban, E. & Davidson, S. DeltaGrad: rapid retraining of machine learning models. In *Proc. 37th International Conference on Machine Learning* Vol. 119, 10355–10366 (PMLR, 2020).
56. Izzo, Z., Smart, M. A., Chaudhuri, K. & Zou, J. Approximate data deletion from machine learning models. In *Proc. 24th International Conference on Artificial Intelligence and Statistics* Vol. 130, 2008–2016 (PMLR, 2021).
57. Everitt, T. & Hutter, M. Avoiding wireheading with value reinforcement learning. In *Artificial General Intelligence. AGI 2016. Lecture Notes in Computer Science* (eds Steunebrink, B. et al.) Vol. 9782 (Springer, 2016).
58. Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J. & Garrabrant, S. Risks from learned optimization in advanced machine learning systems. Preprint at https://doi.org/10.48550/arXiv.1906.01820 (2019).
59. Pistono, F. & Yampolskiy, R. V. Unethical research: how to create a malevolent artificial intelligence. Preprint at https://doi.org/10.48550/arXiv.1605.02817 (2016).
60. Picardi, C., Paterson, C., Hawkins, R., Calinescu, R. & Habli, I. Assurance argument patterns and processes for machine learning in safety-related systems. In *Proc. Workshop on Artificial Intelligence Safety (SafeAI 2020). CEUR Workshop Proceedings* 23–30 (CEUR, 2020).
61. Wabersich, K. J., Hewing, L., Carron, A. & Zeilinger, M. N. Probabilistic model predictive safety certification for learning-based control. *IEEE Trans. Automat. Contr.* **67**, 176–188 (2022).
62. Wen, M. & Topcu, U. Constrained cross-entropy method for safe reinforcement learning. *IEEE Trans. Automat. Contr.* **66**, 3123–3137 (2021).
63. Zanella-Béguelin, S. et al. Analyzing information leakage of updates to natural language models. In *CCS '20: Proc. 2020 ACM SIGSAC Conference on Computer and Communications Security* 363–375 (ACM, 2020).
64. Wang, Z., Chen, C. & Dong, D. A dirichlet process mixture of robust task models for scalable lifelong reinforcement learning. *IEEE Trans. Cybern.* **1**, 12 (2022).
65. Zou, A. et al. Universal and transferable adversarial attacks on aligned language models. Preprint at https://doi.org/10.48550/arXiv.2307.15043 (2023).
66. Ilyas, A. et al. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems* Vol. 32 (NeurIPS, 2019).
67. He, R. D., Han, Z. Y., Yang, Y. & Yin, Y. L. Not all parameters should be treated equally: deep safe semi-supervised learning under class distribution mismatch. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 36, 6874–6883 (AAAI, 2022).
68. Aghakhani, H., Meng, D., Wang, Y.-X., Kruegel, C. & Vigna, G. Bullseye polytope: a scalable clean-label poisoning attack with improved transferability. In *Proc. IEEE Symposium on Security & Privacy* 159–178 (IEEE, 2021).
69. Liu, Y. et al. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022 IEEE Conference on Computer Communications* 280–289 (IEEE, 2022).
70. Meinke, A. & Hein, M. Towards neural networks that provably know when they don't know. In *International Conference on Learning Representations* (ICLR, 2020).
71. Abdelfattah, S., Kasmarik, K. & Hu, J. A robust policy bootstrapping algorithm for multi-objective reinforcement learning in non-stationary environments. *Adapt. Behav.* **28**, 273–292 (2020).
72. Djeumou, F., Cubuktepe, M., Lennon, C. & Topcu, U. Task-guided inverse reinforcement learning under partial information. In *Proc. 32nd International Conference on Automated Planning and Scheduling* (ICAPS, 2021).
73. Ring, M. & Orseau, L. Delusion, survival, and intelligent agents. In *Artificial General Intelligence Lecture Notes in Computer Science* (eds Schmidhuber, J. et al.) 11–20 (Springer, 2011).
74. Soares, N., Fallenstein, B., Yudkowsky, E. & Armstrong, S. Corrigibility. In *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop* (AAAI, 2015).
75. Yampolskiy, R. V. Leakproofing the singularity: artificial intelligence confinement problem. *J. Conscious. Stud.* **19**, 194–214 (2012).
76. Yampolskiy, R. V. in *Philosophy and Theory of Artificial Intelligence Studies in Applied Philosophy, Epistemology and Rational Ethics* (ed. Müller, V. C.) 389–396 (Springer, 2013).
77. Yampolskiy, R. V. Taxonomy of pathways to dangerous AI. Preprint at https://doi.org/10.48550/arXiv.1511.03246 (2015).
78. Wheatley, S., Sovacool, B. K. & Sornette, D. Reassessing the safety of nuclear power. *Energy Res. Soc. Sci.* **15**, 96–100 (2016).
79. Kobayashi, T. et al. (eds) Formal modelling of safety architecture for responsibility-aware autonomous vehicle via event-B refinement. In *Formal Methods, FM 2023 Lecture Notes in Computer Science* Vol. 14000 (eds Katoen, J. P. et al.) 533–549 (2023).
80. Tay, E. B., Gan, O. P. & Ho, W. K. A study on real-time artificial intelligence. *IFAC Proc. Vol.* **30**, 109–114 (1997).
81. Hibbard, B., Bach, J., Goertzel, B. & Iklé, M. Avoiding unintended AI behaviors. In *Artificial General Intelligence: Lecture Notes in Computer Science* (eds Bach, J. et al.) 107–116 (Springer, 2012).
82. Sezener, C. E. Bieger, J., Goertzel, B. & Potapov, A. Inferring human values for safe AGI design. In *Artificial General Intelligence* (eds Bieger, J. et al.) Vol. 9205, 152–155 (AGI, 2015).
83. El Mhamdi, E. & Guerraoui, R. When neurons fail. In *IEEE International Parallel and Distributed Processing Symposium* 1028–1037 (IEEE, 2017).
84. Freiesleben, T. & Grote, T. Beyond generalization: a theory of robustness in machine learning. *Synthese* **202**, 109 (2023).
85. Sanneman, L. & Shah, J. Transparent value alignment. In *HRI '23: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* 557–560 (ACM, 2023).
86. Abbeel, P. & Ng, A. Y. Apprenticeship learning via inverse reinforcement learning In *ICML '04: Proc. 21st International Conference on Machine Learning* (ACM, 2004).
87. Murphy, B. et al. Learning effective and interpretable semantic models using non-negative sparse embedding.In *Proc. COLING 2012* 1933–1950 (The COLING 2012 Organizing Committee, 2012).
88. Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T. & Hovy, E. SPINE: sparse interpretable neural embeddings. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 32 (2017).
89. Shaham, U., Yamada, Y. & Negahban, S. Understanding adversarial training: increasing local stability of supervised models through robust optimization. *Neurocomputing* **307**, 195–204 (2018).
90. Wu, G., Hashemi, M. & Srinivasa, C. PUMA: performance unchanged model augmentation for training data removal. In *Proc. 36th AAAI Conference on Artificial Intelligence* (AAAI, 2022).
91. Jing, S. & Yang, L. A robust extreme learning machine framework for uncertain data classification. *J. Supercomput.* **76**, 2390–2416 (2020).
92. Gan, H. T., Li, Z. H., Fan, Y. L. & Luo, Z. Z. Dual learning-based safe semi-supervised learning. *IEEE Access* **6**, 2615–2621 (2018).

93. Engstrom, L. et al. Adversarial robustness as a prior for learned representations. Preprint at https://doi.org/10.48550/arXiv.1906.00945 (2019).
94. Brophy, J. & Lowd, D. Machine unlearning for random forests. In *Proc. 38th International Conference on Machine Learning* Vol. 139, 1092–1104 (PMLR, 221).
95. Chundawat, V. S., Tarun, A. K., Mandal, M. & Kankanhalli, M. Zero-shot machine unlearning. *IEEE Trans. Inf. Forensics Secur.* **18**, 2345–2354 (2023).
96. Chen, J. et al. ATOM: robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track: Lecture Notes in Computer Science* (eds Oliver, N. et al.) 430–445 (Springer, 2021).
97. Lakkaraju, H., Kamar, E., Caruana, R. & Horvitz, E. Identifying unknown unknowns in the open world: representations and policies for guided exploration. In *AAAI'17: Proc. 31st AAAI Conference on Artificial Intelligence* 2124–2132 (ACM, 2017).
98. Zhuo, J. B., Wang, S. H., Zhang, W. G. & Huang, Q. M. Deep unsupervised convolutional domain adaptation. In *MM '17: Proc. 25th ACM International Conference on Multimedia* 261–269 (ACM, 2017).
99. Bossens, D. M. & Bishop, N. Explicit explore, exploit, or escape (E4): near-optimal safety-constrained reinforcement learning in polynomial time. *Mach. Learn.* **112**, 817–858 (2023).
100. Massiani, P. F., Heim, S., Solowjow, F. & Trimpe, S. Safe value functions. *IEEE Trans. Automat. Contr.* **68**, 2743–2757 (2023).
101. Shi, M., Liang, Y. & Shroff, N. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. In *ICML'23: Proc. 40th International Conference on Machine Learning* Article 1291, 31243–31268 (ACM, 2023).
102. Hunt, N. et al. Verifiably safe exploration for end-to-end reinforcement learning. In *HSCC '21: Proc. 24th International Conference on Hybrid Systems: Computation and Control* (ACM, 2021).
103. Ma, Y. J., Shen, A., Bastani, O. & Dinesh, J. Conservative and adaptive penalty for model-based safe reinforcement learning. In *Proc. AAAI Conference on Artificial Intelligence* **36**, 5404–5412 (AAAI, 2022).
104. Zwane, S. et al. Safe trajectory sampling in model-based reinforcement learning. In *19th International Conference on Automation Science and Engineering (CASE)* (IEEE, 2023).
105. Fischer, J., Eyberg, C., Werling, M. & Lauer, M. Sampling-based inverse reinforcement learning algorithms with safety constraints. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* 791–798 (IEEE, 2021).
106. Zhou, Z., Liu, G. & Zhou, M. A robust mean-field actor-critic reinforcement learning against adversarial perturbations on agent states. *IEEE Trans. Neural Netw. Learn. Syst.* 1–12 (2023).
107. Bazzan, A. L. C. Aligning individual and collective welfare in complex socio-technical systems by combining metaheuristics and reinforcement learning. *Eng. Appl. Artif. Intell.* **79**, 23–33 (2019).
108. Christoffersen, P. J., Haupt, A. A. & Hadfield-Menell, D. Get it in writing: formal contracts mitigate social dilemmas in multi-agent RL. In *AAMAS '23: Proc. 2023 International Conference on Autonomous Agents and Multiagent Systems* 448–456 (ACM, 2023).
109. Christiano, P. F. et al. Deep reinforcement learning from human preferences. In *NIPS'17: Proc. 31st International Conference on Neural Information Processing Systems* 4302–4310 (2017).
110. Kaushik, D., Hovy, E. & Lipton, Z. C. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations* (ICLR, 2020).
111. Li, ZY., Zeng, J., Thirugnanam, A. & Sreenath, K. Bridging model-based safety and model-free reinforcement learning through system identification of low dimensional linear models. In *Robotics Science and Systems* Paper 033(2022).
112. Zhu, X., Kang, S. C. & Chen, J. Y. A contact-safe reinforcement learning framework for contact-rich robot manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 2476–2482 (IEEE, 2022).
113. Terra, A., Riaz, H., Raizer, K., Hata, A. & Inam, R. Safety vs. efficiency: AI-based risk mitigation in collaborative robotics. In *6th International Conference on Control, Automation and Robotics*151–160 (ICCAR, 2020).
114. Adebayo, J. et al. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* Vol. 31 (NeurIPS, 2018).
115. Carlini, N. & Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)* 39–57 (IEEE, 2017).
116. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 427–436 (IEEE, 2015).
117. Kaufmann, M. et al. Testing robustness against unforeseen adversaries. Preprint at https://doi.org/10.48550/arXiv.1908.08016 (2019).
118. Ray, A., Achiam, J. & Amodei, D. Benchmarking safe exploration in deep reinforcement learning. *OpenAI* https://openai.com/index/benchmarking-safe-exploration-in-deep-reinforcement-learning/ (2023).
119. Hendrycks, D. et al. What would Jiminy Cricket do? Towards agents that behave morally. In *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks 1* (NeurIPS, 2021).
120. Gardner, M. et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (eds Cohn, T. et al.) 1307–1323 (Association for Computational Linguistics, 2020).
121. Spears, D. F. Agent technology from a formal perspective. In *NASA Monographs in Systems and Software Engineering* (eds Rouff, C. A. et al.) 227–257 (Springer, 2006).
122. Wozniak, E., Cârlan, C., Acar-Celik, E. & Putzer, H. A safety case pattern for systems with machine learning components. In *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops* Vol. 12235, 370–382 (Springer, 2020).
123. Hendrycks, D. & Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. International Conference on Learning Representations* (ICLR, 2019).
124. Gholampour, P. & Verma, R. Adversarial robustness of phishing email detection models. In *IWSPA '23: Proc. 9th ACM International Workshop on Security and Privacy Analytics* 67–76 (2023).
125. Nanda, N., Chan, L., Lieberum, T., Smith, J. & Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *11th International Conference on Learning Representations* (ICLR, 2023).
126. Olsson, C. et al. In-context learning and induction heads. *Transformer Circuits Thread* https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html (2022).
127. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations* Poster (ICLR, 2015).
128. Kim, B. et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In *Proc. 35th International Conference on Machine Learning* 2668–2677 (PMLR, 2018).

129. Karpathy, A., Johnson, J. & Fei-Fei, L. Visualizing and understanding recurrent networks. Preprint at https://doi.org/10.48550/arXiv.1506.02078 (2016).
130. Morcos, A. S., Barrett, D. G. T., Rabinowitz, N. C. & Botvinick, M. On the importance of single directions for generalization. Preprint at https://doi.org/10.48550/arXiv.1803.06959 (2018).
131. Gyevnar, B., Wang, C., Lucas, C. G., Cohen, S. B. & Albrecht, S. V. Causal explanations for sequential decision-making in multi-agent systems. In *AAMAS '24: Proce. 23rd International Conference on Autonomous Agents and Multiagent Systems* 771–779 (ACM, 2024).
132. Okawa, Y., Sasaki, T. & Iwane, H. Automatic exploration process adjustment for safe reinforcement learning with joint chance constraint satisfaction. *IFAC-PapersOnLine* **53**, 1588–1595 (2020).
133. Gan, H. T., Luo, Z. Z., Meng, M., Ma, Y. L. & She, Q. S. A risk degree-based safe semi-supervised learning algorithm. *Int. J. Mach. Learn. Cybern.* **7**, 85–94 (2016).
134. Zhang, Y. X. et al. Barrier Lyapunov Function-based safe reinforcement learning for autonomous vehicles with optimized backstepping. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 2066–2080 (2024).
135. Nicolae, M.-I., Sebban, M., Habrard, A., Gaussier, E. & Amini, M.-R. Algorithmic robustness for semi-supervised (ϵ, γ, τ)-good metric learning. In *International Conference on Neural Information Processing* (ICONIP, 2015).
136. Khan, W. U. & Seto, E. A 'do no harm' novel safety checklist and research approach to determine whether to launch an artificial intelligence-based medical technology: introducing the biological-psychological, economic, and social (BPES) framework. *J. Med. Internet Res.* **25**, e43386 (2023).
137. Schumeg, B., Marotta, F. & Werner, B. Proposed V-model for verification, validation, and safety activities for artificial intelligence. In *2023 IEEE International Conference on Assured Autonomy (ICAA)* 61–66 (IEEE, 2023).
138. Kamm, S., Sahlab, N., Jazdi, N. & Weyrich, M. A concept for dynamic and robust machine learning with context modeling for heterogeneous manufacturing data. *Procedia CIRP* **118**, 354–359 (2023).
139. Costa, E., Rebello, C., Fontana, M., Schnitman, L. & Nogueira, I. A robust learning methodology for uncertainty-aware scientific machine learning models. *Mathematics* **11**, 74 (2023).
140. Aksjonov, A. & Kyrki, V. A Safety-critical decision-making and control framework combining machine-learning-based and rule-based algorithms. *SAE Int. J. Veh. Dyn. Stab. NVH* **7**, 287–299 (2023).
141. Antikainen, J. et al. A deployment model to extend ethically aligned AI implementation method ECCOLA. In *29th IEEE International Requirements Engineering Conference Workshops* (eds Yue, T. & Mirakhorli, M.) 230–235 (IEEE, 2021).
142. Vakkuri, V., Kemell, K. K., Jantunen, M., Halme, E. & Abrahamsson, P. ECCOLA — a method for implementing ethically aligned AI systems. *J. Syst. Softw.* **182**,111067 (2021).
143. Zhang, H., Shahbazi, N., Chu, X. & Asudeh, A. FairRover: explorative model building for fair and responsible machine learning. In *DEEM '21: Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning* Article 5, 1–10 (ACM, 2021).
144. Coston, A. et al. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* 690–704 (IEEE, 2023).
145. Gittens, A., Yener, B. & Yung, M. An adversarial perspective on accuracy, robustness, fairness, and privacy: multilateral-tradeoffs in trustworthy ML. *IEEE Access* **10**, 120850–120865 (2022).
146. Taylor, J., Yudkowsky, E., LaVictoire, P. & Critch, A. Alignment for advanced machine learning systems. In *Ethics of Artificial Intelligence* (ed. Liao, S. M.) 342–382 (Oxford Academic, 2016).
147. Sotala, K. & Yampolskiy, R. V. Responses to catastrophic AGI risk: a survey. *Phys. Scr.* **90**, 018001 (2014).
148. Johnson, B. Metacognition for artificial intelligence system safety — an approach to safe and desired behavior. *Saf. Sci.* **151**, 105743 (2022).
149. Hatherall, L. et al. Responsible agency through answerability: cultivating the moral ecology of trustworthy autonomous systems. In *TAS '23: Proc. First International Symposium on Trustworthy Autonomous Systems* 50, 1–5 (2023).
150. Stahl, B. C. Embedding responsibility in intelligent systems: From AI ethics to responsible AI ecosystems. *Sci. Rep.* **13**, 7586 (2023).
151. Samarasinghe, D. Counterfactual learning in enhancing resilience in autonomous agent systems. *Fron. Artif. Intell.* **6**, 1212336 (2023).
152. Diemert, S., Millet, L., Groves, J. & Joyce, J. Safety integrity levels for artificial intelligence. In *Computer Safety, Reliability, and Security. SAFECOMP 2023 Workshops* (eds Guiochet, J. et al.) Vol. 14182, 397–409 (Springer, 2023).
153. Wang, J. & Jia, R. Data Banzhaf: a robust data valuation framework for machine learning. In *Proc. 26th International Conference on Artificial Intelligence and Statistics (AISTATS)* Vol. 206, 6388–6421 (PMLR, 2023).
154. Everitt, T., Filan, D., Daswani, M. & Hutter, M. Self-modification of policy and utility function in rational agents. In *Artificial General Intelligence: 9th International Conference, AGI 2016* (eds Steunebrink, B. et al.) Vol 9782 (Springer, 2016).
155. Badea, C. & Artus, G. Morality, machines, and the interpretation problem: a value-based, wittgensteinian approach to building moral agents. In *Artificial Intelligence, AI 2022* (eds Bramer, M. & Stahl, F.) Vol. 39, 124–137 (2022).
156. Umbrello, S. Beneficial artificial intelligence coordination by means of a value sensitive design approach. *Big Data Cogn. Comput* **3**, 5 (2019).
157. Yampolskiy, R. & Fox, J. Safety engineering for artificial general intelligence. *Topoi* **32**, 217–226 (2012).
158. Weld, D. et al. (eds) The first law of robotics. In *Safety and Security in Multiagent Systems: Lecture Notes in Computer Science* (eds Barley, M. et al.) 90–100 (Springer, 2009).
159. Matthias, A. The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **6**, 175–183 (2004).
160. Farina, L. Artificial intelligence systems, responsibility and agential self-awareness. In *Philosophy and Theory of Artificial Intelligence 2021* Vol. 63 (ed. Muller, V. C.) 15–25 (2022).
161. Lee, A. T. *Flight Simulation: Virtual Environments in Aviation* (Routledge, 2017).
162. Torres, E. P. *Human Extinction: A History of the Science and Ethics of Annihilation* (Routledge, 2023).
163. Bostrom, N. Existential risks: analyzing human extinction scenarios and related hazards. *J. Evol. Technol.* **9**, 1–30 (2002).
164. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford Univ. Press, 2014).
165. Ord, T. *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books, 2020).
166. Dafoe, A. & Russell, S. Yes, we are worried about the existential risk of artificial intelligence. *MIT Technology Review* https://www.technologyreview.com/2016/11/02/156285/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/ (2016).
167. Bronson, R. Measuring existential risk. *Peace Policy* https://peacepolicy.nd.edu/2023/12/06/measuring-existential-risk/ (2023).

168. Roose, K. A.I. poses 'risk of extinction', industry leaders warn. *The New York Times* https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html (2023).
169. Statement on AI risk. *Center for AI Safety* https://www.safe.ai/work/statement-on-ai-risk (2023).
170. Weidinger, L. et al. Sociotechnical safety evaluation of generative ai systems. Preprint at https://doi.org/10.48550/arXiv.2310.11986 (2023).
171. Anwar, U. et al. Foundational challenges in assuring alignment and safety of large language models. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=oVTkOs8Pka (2024).
172. Ganin, Y. et al. in *Domain Adaptation in Computer Vision Applications* (ed. Csurka, G.) 189–209 (Springer, 2017).
173. Balaji, Y., Sankaranarayanan, S. & Chellappa, R. MetaReg: towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems* Vol. 31 (eds Bengio, S. et al.) (NeurIPS, 2018).
174. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks. Preprint at https://doi.org/10.48550/arXiv.1706.06083 (2019).
175. Turchetta, M. et al. Safe exploration for interactive machine learning. In *Advances in Neural Information Processing Systems* Vol. 32 (eds Wallach, H. et al.) (NeurIPS, 2019).
176. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Proc. Syst.* **35**, 27730–27744 (2022).
177. Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of- distribution examples in neural networks. In *Proc. 5th International Conference on Learning Representations* (ICLR, 2017).

Acknowledgements

This research was made possible through the generous support of several funding bodies. Primary support came from the AI2050 programme at Schmidt Sciences (grant 24-66924), which supported A.K.'s contributions. Additional funding was provided by the UKRI Arts and Humanities Research Council (grant AH/X007146/1). B.G.'s work was supported by the UKRI Centre for Doctoral Training in Natural Language Processing through a joint grant (EP/S022481/1) from UK Research and Innovation and the University of Edinburgh's School of Informatics and School of Philosophy, Psychology and Language Sciences.

Author contributions

A.K. conceived and conceptualized the study, guided the literature selection and review methodology, supervised the research direction, and interpreted both qualitative and quantitative analyses. B.G. conducted the systematic literature review, performed data analysis and visualization, and documented the findings. Both authors wrote, reviewed and approved the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Atoosa Kasirzadeh.

Peer review information *Nature Machine Intelligence* thanks Fabian Ferrari, Sebastian Porsdam Mann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.