Humanities & Social Sciences Communications

REVIEW ARTICLE

https://doi.org/10.1057/s41599-025-04814-y OPEN

Focal points and blind spots of human-centered AI: AI risks in written online media

Marcell Sebestyén [1]✉

There is a strong tendency in prevailing discussions about artificial intelligence (AI) to focus predominantly on human-centered concerns, thereby neglecting the broader impacts of this technology. This paper presents a categorization of AI risks highlighted in public discourse, as reflected in written online media accounts, to provide a background for its primary focus: exploring the dimensions of AI threats that receive insufficient attention. Particular emphasis is dedicated to the ignored issues of animal welfare and the psychological impacts on humans, the latter of which surprisingly remains inadequately addressed despite the prevalent anthropocentric perspective of the public conversation. Moreover, this work also considers other underexplored dangers of AI development for the environment and, hypothetically, for sentient AI. The methodology of this study is grounded in a manual selection and meticulous, thematic, and discourse analytical manual examination of online articles published in the aftermath of the AI surge following ChatGPT's launch in late 2022. This qualitative approach is specifically designed to overcome the limitations of automated, surface-level evaluations typically used in media reviews, aiming to provide insights and nuances often missed by the mechanistic and algorithm-driven methods prevalent in contemporary research. Through this detail-oriented investigation, a categorization of the dominant themes in the discourse on AI hazards was developed to identify its overlooked aspects. Stemming from this evaluation, the paper argues for expanding risk assessment frameworks in public thinking to a morally more inclusive approach. It calls for a more comprehensive acknowledgment of the potential harm of AI technology's progress to non-human animals, the environment, and, more theoretically, artificial agents possibly attaining sentience. Furthermore, it calls for a more balanced allocation of focus among prospective menaces for humans, prioritizing psychological consequences, thereby offering a more sophisticated and capable strategy for tackling the diverse spectrum of perils presented by AI.

[1] Department of Philosophy and History of Science, Faculty of Economic and Social Sciences, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary. Link to institutional website: https://www.filozofia.bme.hu/people/marcell.sebestyen?lang=EN. ✉email: marcell.sebestyen@filozofia.bme.hu; marcell.sebestyen@gmail.com

Introduction

This paper examines the ongoing public discourse on the risks associated with AI, as reflected in written online media coverage. Specifically, a classification framework of AI threats consisting of 37 + 1 categories is introduced, derived from a thematic and discourse analysis of how these dangers are portrayed in online media articles. The key purpose of this study is to reveal that the current discussion surrounding AI threats overlooks multiple critical areas: the psychological effects of AI on humans, the dangers posed to non-human animals (referred to simply as animals), the environment, and artificial agents potentially evolving into self-aware and/or sentient entities due to the development of the technology.

The structure of this paper is as follows: subsequent to this introductory section, which aims to illuminate the fundamental ideas to be elaborated on later, the second section presents the theoretical background of the investigation, focusing on the thematic and discourse analysis of the written online media coverage of AI risks, and derived from that, the third section outlines the categorization of these threats. The fourth section delves into the evaluation of the findings, conveying the principal aim of this study: identifying blind spots within the public discourse and highlighting the necessity to adjust and broaden its focus beyond exclusively human concerns. Within the domain of anthropocentric perspectives, it is argued that attention must be redirected toward specific elements, particularly the psychological implications. Additionally, it will be maintained that a more inclusive approach to risk assessment is crucial, considering the interests of non-human entities—foremost among these, a vast range of animals—and recognizing them as subjects of moral concern due to their capacity for suffering. The concluding fifth section will outline the findings and propose a significant shift in discussing AI perils, advocating for a more comprehensive framework that thoroughly addresses the diverse threats posed by AI advancements.

Significance and challenges of AI risk assessment. Many hold the view that the survival of living organisms, as well as the quality of life that the Earth offers to beings living on it, are of fundamental importance. Therefore, it appears to be an essential task to consider and seek to prevent any circumstances that could possibly threaten the continuation of the existence of the natural world, diminish the living conditions on our planet, or cause suffering to any sentient entities. This holds true even though a significant portion of society seems to underestimate the cognitive abilities of animals compared to scientific evidence (Leach et al. 2023). Nevertheless, the recognition of sentience across a broad spectrum of animals is increasingly reflected in legal and cultural frameworks. (Treaty of Lisbon 2007, 49; Animal Welfare Sentience Act (2022), Andrews et al. (2024))

These apprehensions arise in relation to every novel, highly potent technology invented by humanity. However, with advancements like AI, which has an immense potential to become extremely powerful and versatile and is already altering the way we live (Salvi and Singh 2023, pp. 5441–43), these concerns are particularly well-founded and relevant.

In order to effectively address the potential risks posed by AI, it is first necessary to ascertain the nature of the various dangers we might encounter, a task that is inherently challenging to accomplish. Merely pondering upon the expression 'superintelligent AI,' which denotes an artificial agent possessing a level of cognitive capability that considerably surpasses human intellect, can quickly lead to the conclusion that attempts to foresee the approaching adversities are futile. With the mental faculties of Homo sapiens, it is—by the very meaning of the term *superintelligent*—unattainable to anticipate all the actions such an entity might undertake (Bostrom 2014, p. 52). Nevertheless, given the immense scale of the stakes, efforts must be made to predict the scenarios that might unfold.

Some hazards involved in this conversation are notably easier to formulate prognoses about. Besides frequent discussions about employment shifts as a result of a new wave of automatization and its possible consequences, the perils posed by AI seem to be overly emphasized regarding artificial agents potentially leading to annihilation or enslavement of humanity, both in the scientific literature, as well as in everyday narratives (Turchin and Denkenberger 2020, pp. 147–48). Clearly, these topics draw greater attention than their more down-to-earth counterparts, but be that as it may, it appears evident that additional steps should be taken to reveal the less severe but intuitively more realistic and, fortunately, potentially also more predictable outcomes of AI technology.

Need for expanding and refocusing the AI risk framework. There seems to be a middle ground between the far-fetched, extremely severe, even catastrophic scenarios threatening the very existence of human civilization and the rather obvious fears, such as automation-driven job losses. A much stronger emphasis must be put on the dangers that fall into this zone of insufficient focus across both academic research and public opinion. The issues of psychological damage inflicted on humans, along with the technology's sustainability concerns, undoubtedly belong to this area. Yet, an even more alarming oversight is the almost complete neglect of the suffering caused to animals by AI, which demands urgent attention.

Various and significant risks are posed to the human psyche, including but not limited to eroding mental health and emotional well-being through manipulative relationships and deceptive content, fostering digital addiction, social isolation, as well as diminished human functions (Shanmugasundaram and Tamilarasu 2023; Ienca 2023). In the case of animals, primary concerns that stand out as particularly pressing and demand urgent attention emerge from AI-driven enhancements in factory farming efficiency, potentially exacerbating already appalling conditions for animals, alongside algorithmic bias against animals that may be capable of solidifying the exploitation of animals in the social fabric (Singer and Tse 2023, pp. 541–547). Concerning the ecological effect and the feasibility, the associated perils include energy consumption and, in particular, the greenhouse gas emissions it induces, the technology's water footprint, and the demand for specific materials such as lithium or cobalt, to name a few, across the life cycle of an AI system (Ligozat et al. 2022).

This study seeks to draw attention to the insufficiently addressed or utterly disregarded menaces and stresses the importance of prioritizing the highly realistic perils among these. Having said all this, with each new development that we witness and are likely to see in the near future, it becomes increasingly more challenging to determine which ideas hide genuine threats and which should still be considered unfounded speculations. Therefore, even the possibilities that might strike one as extremely unlikely, bearing in mind the tremendous risks they carry, must be taken into account to some degree, even if we pay closer attention to more probable eventualities. For instance, Bostrom argues that the possible amount and severity of suffering that artificial agents might have to endure in the future is so monumental in extent that it by far exceeds the aggregated agony of all biological organisms that have ever inhabited our planet (Bostrom 2014, pp. 101–103). Despite the highly speculative nature and extremely low likelihood of these scenarios, the immense stakes, often referred to as 'astronomical suffering',

'mind crime', and especially the combination of the two (Bostrom 2014, p. 152; Gloor and Althaus 2016; Sotala and Gloor 2017), provide valid grounds for their inclusion in our analysis. Nonetheless, this work argues that greater focus should be directed toward developments affecting human psychology, animals, and the environment.

Exploring the ongoing discourse from a different angle, it must be pointed out that the public discussion, as well as the ethical debate surrounding the dangers originating from AI is predominantly human-centered in the sense that the inquiries and reports on the topic tend to be fixating on the impact of the technology *on humans exclusively* (Owe and Baum 2021; Rigley et al. 2023, pp. 844–848). One flaw of this perspective is that it overlooks the fact that mankind constitutes only an insignificant proportion of the total animal population on Earth, not to mention other forms of biological life.

Risk assessment must factor in the interests of entities capable of experiencing subjective sensations comprising joy and suffering. For a substantial segment of the animal kingdom, the capacity for pain perception is clearly established, encompassing mammals, birds (Low 2012), and arguably, fish (Balcombe 2017, pp. 71–85; Braithwaite 2010). This investigation also extends the scope from biological beings to include other dimensions. Specifically, it addresses, though with less emphasis, the speculative issue of artificial agents that might have the potential to reach a state of sentience (James and Scott 2008). Adopting Bentham's stance, sentience is considered a necessary and sufficient prerequisite for agents to have interests and, thereby, also a satisfactory condition for possessing some kind of moral status (Bortolotti et al. 2013). Additionally, the concept of the natural environment as a moral subject will be addressed, a notion that aligns with several ethical frameworks (Palmer, McShane, and Sandler 2014).

Analysis of the media coverage of AI risks

This section will introduce the methodology for identifying focal points and blind spots—with the latter serving as the central focus of this study—in the public perception of AI risks, examined through the online media that both shapes and represents public opinion.

Written online media as public reflection. The press serves as a paradoxical medium, with journalists both representing and actively shaping public opinion and collective societal attitudes. (McLuhan 1994, p. 213)

Since, for the general public, one of the primary sources of information today is online news coverage, the issues addressed in digital media strongly shape people's opinions and bring certain areas into focus (Zhou and Moy 2007, pp. 81–84; Shrum 2017, pp. 9–10; Sun et al. 2020, p. 1). The topic of AI is no exception to this influence (Chuan et al. 2019, p. 339).

Accordingly, through their cultural influence, journalists have an impact on the trajectory along which these tendencies—in this particular instance, the integration of AI into the fabric of society—unfold. The relation is two-sided, though, in the sense that journalists also represent society through their personal news decisions (Patterson and Donsbagh 1996), therefore, the questions they discuss reflect the concerns and topics of everyday people.

Complementing this, Neri and Cozman (2020) demonstrated that experts in the spotlight often drive AI risk perception, playing pivotal roles in shaping and moderating the discourse, particularly through social media platforms. Be that as it may, this paper will not deeply delve into this phenomenon nor the complex dynamics between laypeople and journalists. Instead, aligned with the previous paragraph's assertions, it will be assumed that columnists either represent broader societal views or direct public attention to prevalent issues and, as a result, their narratives reflect or converge on the most frequently discussed and most captivating or unsettling matters among the general public. According to this, this paper starts from the premise that an exploration of written online media coverage will uncover meaningful understandings of public opinion. This analysis extends beyond established digital news outlets to include a variety of online media sources, as non-traditional media forums serve a comparable function and mirror a similar format.

Limitations of prior research on AI risks in media. Research has already explored how AI technology and the threats it poses are portrayed in the news media, specifically in the studies conducted by Chuan et al. (2019), Sun et al. (2020), Nguyen and Hekman (2022), and Nguyen (2023). Nevertheless, their work concentrated on a longer timeframe that preceded the recent AI surge, which was undoubtedly triggered by the launch of the large language model (LLM), ChatGPT, at the end of 2022, significantly enhancing public awareness of generative AI technologies (Waters 2023; Roe and Perkins 2023, p. 2). Accordingly, the referenced 'pre-ChatGPT' media analyses could not capture the most recent perspectives in the discourse solely because of the timeframe of their investigation. Furthermore, while the authors' findings on the prevalence of topics such as bias, surveillance, job losses, and cyberattacks in public dialog are consistent with my own results, my greater emphasis on the dimension of risks and the consequent more comprehensive exploration of AI threats provide deeper and novel insights into the matter, which are also more current due to the later date range.

It is also important to acknowledge the research conducted by Xian et al. (2024), which explored news articles from a timeframe partly overlapping with the scope of this study. Nonetheless, their attention to the aspect of dangers surrounding AI was comparatively less substantial, therefore, this paper provides a more complex understanding of this aspect.

Overview of the methodology. A summary of the procedure followed is provided in Fig. 1 to promote a quick and transparent overview of the method used.

In the first phase, articles were collected by conducting an online search using predefined keywords related to AI and risks. A substantial number of articles were found, but only those presenting general discussions of AI threats were selected, and those focusing on specific aspects were excluded.

As a typical example, the publication "*15 AI risks businesses must confront and how to address them*" (Pratt 2024) was not selected due to its narrow focus, specifically on business-related concerns, while "*The 15 Biggest Risks Of Artificial Intelligence*" (Marr 2023), was chosen for its broader perspective.

This process yielded 56 online media articles published between November 2022 and October 2024, all offering an overview of the dangers of AI.

Preliminary experiments were carried out to assess the capability of automated text analysis to identify the risks addressed by the authors in the articles and then to structure them into categories—however, these approaches failed to yield the expected results in terms of comprehensiveness and precise sorting of the perils along the lines of the narratives presented. Consequently, a manual methodology was adopted for further investigation.

A discourse analytical approach was employed to explore the nuanced and context-driven nature of language in texts, which extends beyond the surface meaning of words, considering the

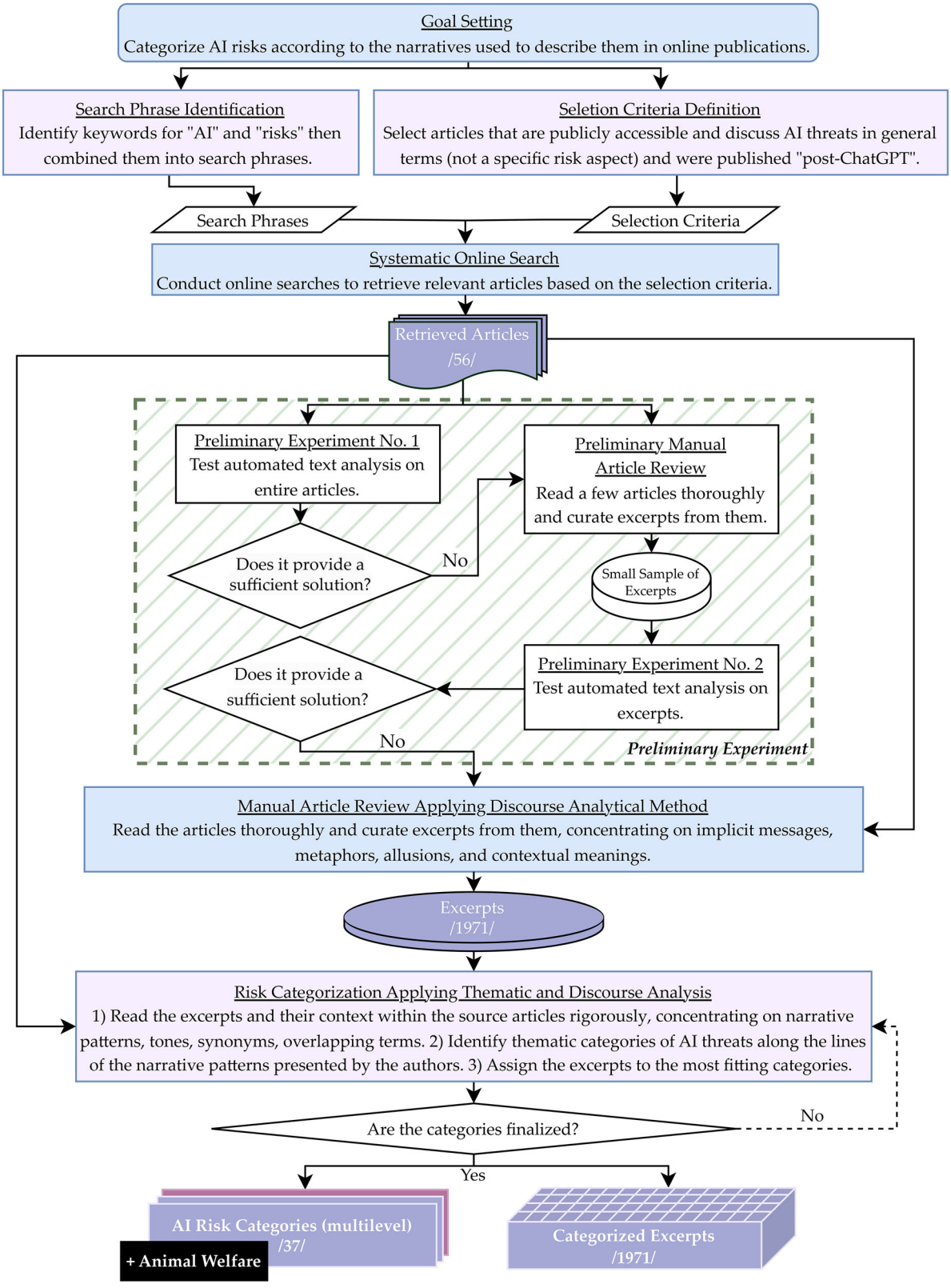

Fig. 1 Flowchart summarizing the media analysis and risk categorization procedure.

broader context of verbal expression, and is focused on uncovering the underlying dynamics that shape the communication of the authors. Discourse refers to the construction of meaning beyond individual statements, addressing the implicit structures and influences behind them, including the authors' motives, background, etc. Guided by discourse analytical interpretation, each article was thoroughly read and critically reviewed, with all mentions of risks or harms, along with the section headings, systematically compiled into two—partially overlapping—lists. As a result of this, 1971 excerpts were

extracted, primarily consisting of quotations, that provided an extensive catalog of the perils covered in online media publications, as well as hundreds of section titles.

As an illustration, the following section headings were extracted from the article "*What Are the Dangers of AI?*" (Reiff 2023): *Deepfakes and Misinformation; Privacy; Job Loss; Bias, Discrimination, and the Issue of "Techno-Solutionism"; Financial Volatility; The Singularity*—as well as excerpts such as '*AI can be used to create and to widely share material that is incorrect*' and '*Deepfakes are an emerging concern*.'

Building on the discourse analytical framing of the data, thematic analysis—integrating elements of framework analysis—was used to identify patterns in the excerpts. The process began with concept-driven coding. That is, from the gathered section headings, an initial set of danger groups (codes) was aggregated to include the most representative ones. As a first step, for each article, the excerpts stemming from that publication were assigned to these risk classes.

For instance, at the outset, a threat category labeled *'Employment'* was established, and the following excerpts from the article "*What Exactly Are the Dangers Posed by A.I.?*" (Metz 2023) were assigned to it: *Job Loss*; *new A.I. could be job killers*; *they could replace some workers*. To provide another example, through discourse-oriented interpretation, the excerpt "*AI recruitment tool … preferred male applicants over females*" (Kaur 2024) was initially framed under the label '*Bias and Discrimination*'. This classification process ensured that categories did not merely reflect explicit labels used in the articles but also captured underlying meanings and evolving discursive constructions.

Afterward, the codes were—over an extended series of successive iterative cycles—refined into categories that captured the media narratives more and more accurately. This process involved splitting, combining, and adjusting the existing groups, as well as, through the application of thematic coding, reassigning the excerpts to the newly created classes at each stage, following repeated readings of the quotations and their context within the original articles.

Illustratively, the original '*Legal*' category—over the course of roughly a dozen iteration phases—was ultimately divided, with most of its elements sorted into the subsets '*Accountability and Liability*' and '*Intellectual Property and Copyright*.'

This process eventually resulted in a multi-level hierarchical classification system consisting of 37 thematic clusters, effectively covering the media narratives to which all the excerpts were assigned.

The following subsections will present a detailed account of the methodology.

Selection of media articles for review. A systematic online search was conducted using predefined keywords related to 'AI' and 'risks' to identify relevant articles for analysis. This approach also mirrors a primary way how the mainstream audience seeks information on the topic online, thereby enabling the identification of articles likely to align with those encountered by the general audience.

Publications were not limited to those from online news portals but also non-conventional media platforms, including opinion sites, blogs, networking sites, as well as academic and institutional websites were considered, taking into account that all the chosen articles serve a corresponding purpose to and share a comparable format with those published by traditional online news media outlets. For the lay audience searching online for publications on AI risks from a wide-ranging perspective, the selected pieces from these sources are just as accessible and likely to be found as those on established digital news channels.

Exclusively, online media articles published after November 2022 were considered, with the latest examined piece from October 2024. This interval is extremely relevant due to the fact that during this period, the hype around AI reached what might well have been its highest peak for years, subsequent to the public launch of the LLM ChatGPT in November 2022, which revealed the capabilities of the technology to the global citizenry and transformed public attitudes toward AI tools, initiating a new phase that calls for fresh insights. Undoubtedly, the current limitations of generative AI were also exposed, but it seems clear from the frenzy during the investigated period that the potential has surpassed the prior expectations within mainstream society, potentially making some of the earlier speculative fears into more tangible realities.

The candidate articles were thoroughly reviewed, and those discussing AI hazards in a broader context rather than concentrating on a single, specific aspect were picked. This decision was grounded in the assumption—which, however, is neither supported nor contradicted by the literature—that these are the pieces the internet readership with a general curiosity about the topic is more likely to discover and read rather than the domain-oriented ones (such as centering on cyber risks, threats for businesses, etc.). Thus, they are regarded as more accurately reflecting what the average reader is likely to have encountered.

Given the manually curated approach employed, evaluating a vast multitude of articles on AI risks published on specific aspects was not feasible, as performing an overarching, non-automated analysis from those publications would have required an unmanageable amount of effort. Fortunately, the number of articles discussing the matter of AI dangers in broad terms—gathered through the online search process—fell in the range that was manageable with the applied methodology.

The exclusive reliance on general-interest articles limits the scope of this study, while topic-specific online media pieces could serve as a valuable source for further research. Despite this, it is assumed that these comprehensive publications provide a clear overview of the most important themes in the public discussion.

Comparison of automated and manual text analysis. A manual analysis was deliberately employed, even if it could only focus on a smaller selection of articles, on the other hand, it enabled a deeper and more nuanced exploration beyond the limitations of automated large-scale dataset analyses. Although the methodological approach of the investigation inherently limits the scope of articles to be examined, algorithmic assessment of a substantial amount of publications reduces the accuracy of the evaluation, overlooks critical details, and results in less nuanced findings, which could not be allowed for the execution of the task that had been set: the categorization of AI risks based on the narratives in which they are presented in online publications.

Research on automated text analysis (Grimmer and Stewart 2013, pp. 268–271; Mahrt and Scharkow 2013, pp. 25–30; Zamith and Lewis 2015, p. 315; Günther and Quandt 2016, p. 86) suggests that while algorithmic evaluations might enhance efficiency and prove useful in many cases, however, they have considerable limitations. Mahrt and Scharkow (2013, p. 29) also indicate that, in some cases, the analysis of a smaller dataset can yield more insightful conclusions than that of a larger one.

Primarily due to the complexity and indeterminacy of human language (Grimmer and Stewart 2013, p. 268; Humphreys and Wang 2018, p. 1277), automated methods might lead to outcomes that are unreliable (Zamith and Lewis 2015, p. 315), practically unverifiable (Grimmer and Stewart 2013, p. 271), prone to misinterpretation (Mahrt and Scharkow 2013, p. 29), or simply are insufficient to perform specific tasks (Günther and Quandt

2016, p. 77)—these concerns must be borne in mind even if they offer extensive sample sizes (Günther and Quandt 2016, p. 86). Consequently, automated techniques cannot replace the layered understanding gained from the close reading of texts and careful reflection, which remain essential for scholarly accuracy (Grimmer and Stewart 2013, p. 268, 270; Günther and Quandt 2016, p. 86; Lind and Meltzer 2021, p. 934).

Programmed text processing generally struggles with deeper contextual and cultural meanings and connotations, including sarcasm, metaphors, as well as complex and figurative rhetorical arguments (Humphreys and Wang 2018, p. 1277). These mechanisms also often fail to recognize narrative patterns and connections between related concepts (Ceran et al. 2015, p. 942).

Furthermore, automated content analysis is often less effective at determining positive or negative tone, which was crucial when dealing with delicately formulated arguments in the selected publications (Conway 2006, p. 196). This distinction was essential to discern what were considered genuine risks from those that were not and were mentioned only as contrasts, coupled with the fact that numerous articles underlined both AI's benefits and perils. Moreover, when using pieces of work from news media platforms as inputs to a model, distinguishing between relevant text and advertisements poses a significant challenge for programmed techniques (Günther and Quandt 2016, p. 77).

The limitations of automated techniques were also indicated by the failure of preliminary experiments with LLM-aided processing of entire news articles, as well as the failure of manually curated excerpts, due to the inability to capture the contextual significance and the overlap of terms describing distinct threats.

Even with this compromise between the breadth of publications and the depth of analysis, the number of hand-selected articles totaled 56, and the overall number of manually chosen and repeatedly evaluated excerpts reached nearly 2000. This dataset was considered to be sufficient to identify narrative patterns in the representation of AI hazards, that is, to identify risk categories *as they thematically appear in the online media*.

Thematic and discourse analysis of media articles. The thematic and discourse analysis of the 56 finally selected online media articles was conducted to identify the various hazards they discuss. In total, over 1971 excerpts were extracted from the publications, comprising mostly short quotations and, in some instances, citations. These items were then systematically categorized into $37 + 1$ thematic clusters. This categorization facilitated the identification of frequently mentioned AI threat areas in online media outlets, those that are seldom referenced—and, crucially, areas that may be overlooked in the media discourse.

The development of the classification framework was an iterative procedure involving repeated reevaluation of excerpts within their full article context and reconsideration of previously proposed categories. This process ultimately resulted in the categories reaching their most detailed resolution, ensuring they could no longer be further disentangled into distinct narratives. It is not implied, however, that some categories could not be slightly further differentiated, but doing so would fail to represent the AI danger narratives *as they appear in the media*.

As already indicated, contrary to the quantitative methods employed in investigations on the news media put forth by Chuan et al. (2019), Nguyen and Hekman (2022), Nguyen (2023), and Xian et al. (2024), this study concentrated on a limited number of articles chosen purposefully to be pieces that deal exactly with the problem of AI risks but are still general-interest writings that are not confined to specialized areas. Moreover, diverging from the approaches mentioned, instead of relying on statistical and sampling methods, topic modeling, and automated content analysis, among other procedures available, a manual qualitative process was chosen, namely the application of thematic and discourse analysis. This decision was motivated by the belief that automated content analysis lacks the capability to discern subtle nuances and fine differences in interpretation, as elaborated in the preceding subsection.

The study's core methodology was rooted in thematic analysis, as it primarily focused on identifying patterns in media narratives rather than developing new theoretical constructs. Discourse analysis was applied mainly in framing the media narratives, guiding the interpretation of these patterns within their broader media and societal context. This means that thematic coding was not merely an inductive categorization process but also guided by an awareness of the rhetorical and discursive structures shaping media AI risk narratives. Additionally, elements of grounded theory contributed to the inductive generation and iterative refinement of thematic codes.

Thematic coding: integrating framework analysis and elements from grounded theory. Since thematic analysis focuses on uncovering and organizing patterns in qualitative data, it provided a structured approach for examining the ways AI risks are addressed in the selected media articles.

Thematic coding (Gibbs 2007, p. 38) was conducted by interpreting each relevant excerpt—whether discussing risks implicitly or explicitly—to extract its contextual meaning, followed by assigning a specific code that linked the passage to the identified idea.

The first set of codes (codebook) emerged after the initial reading of the corpus, aggregating risk categories from the sample's online media articles based on their section headings, functioning as predefined indices—a method known as concept-driven coding. The strategy of compiling a list of thematic ideas and then applying these codes to the text can be described as framework analysis within thematic analysis (Ritchie and Lewis 2003, pp. 220–24; Gibbs 2007, p. 44).

The process of reviewing excerpts within the broader context of their respective articles and systematically indexing them was conducted iteratively. As decisions had to be made on borderline cases between closely related groups regarding which label an excerpt would fall under, definitions of the codes were altered (Gibbs 2007, p. 40).

The process of tagging excerpts with thematic codes extended beyond mere description. It involved integrating these codes into broader categories, which were sometimes later refined or subdivided, and developing analytic codes. Unlike descriptive codes, which closely reflect the authors' explicit expressions, analytic codes provide a deeper understanding by interpreting how the author perceives an issue, drawing on implicit meanings within the text (Gibbs 2007, pp. 42–43).

While this applied method of thematic coding did not fully adopt a grounded theory approach, it incorporated elements of it. Specifically, emergent coding and recursive category refinement were integrated as the process aimed to inductively generate insights in a cyclic process, drawing directly from the data rather than relying on predefined theories, given that the purpose of the investigation was to identify the narratives through which AI risks are presented in the online media. Considering that both the original and iteratively revised codebooks were entirely anchored in the investigated corpus itself—without drawing from any pre-existing scholarly frameworks on AI risks—all ideas reflected in the classification system are 'grounded' in the data, emerging from and supported by it (Gibbs 2007, pp. 49–50).

The comparison of this derived structure with existing classification systems and taxonomies of AI-related challenges in the literature was conducted only later, as will be presented in

sections "Taxonomy and classification of AI risks in the literature" and "Comparison of the proposed categorization with existing frameworks".

Discourse analytical approach: contextual framing and meaning construction. Given that risk communication is inherently language-centered, discourse analysis provided a framework for examining how AI risks are framed and portrayed within the media landscape (Sarangi and Candlin 2003).

The discourse analysis approach examines texts beyond surface meaning, considering the broader context of verbal expression, focusing on uncovering the underlying motives and background, while taking into account extralinguistic elements. (Sarangi and Candlin 2003, p. 116). These might include, in the case of written online media, contextual information, intertextuality, paratextual references, and visual features.

This was achieved by systematically analyzing recurring discursive patterns in AI risk narratives, examining how different risks were framed, and identifying both implicit and explicit rhetorical strategies applied by the authors to construct these danger portrayals. Additionally, intertextual references, such as links to broader societal themes—including automation, governance, and ethical responsibility—were considered to contextualize the positioning of AI threats within public discourse.

In the examination of individual online media articles discussing AI risks, a specific criterion for identifying mentions of risk was applied: the publication must treat the issue as an actual threat rather than simply discussing it in general (not as a hazard) or suggesting it is a danger that is not worth consideration. Mere references to issues without recognizing their significance as dangers did not qualify as a valid mention in this study.

Gee (2014, pp. 80–82) argues that language must be understood in context, as words derive meaning from their application rather than their mere presence. His concept of situated meanings —assembled 'on the spot' based on contextual cues—aligns with the focus employed in this study on examining how terms are used to convey significance.

Given that various keywords correspond to fundamentally distinct issues, the analysis of the mentions carried out in this study went beyond just identifying the presence of terms within an article. Instead, the context in which these keywords were used was minutely examined to determine whether a topic was meaningfully addressed and in what sense the term was applied.

As Gee (2014, pp. 82–85) highlights, language relies on contextual references to convey meaning, reflecting and constructing reality. Therefore, examining terms within their specific contexts is crucial to uncover their significance, focusing on how language shapes meaning in discourse.

Certain matters are solely implied in the articles rather than explicitly stated, as illustrated by the subtle way of referring to existential threats in several cases. For instance—using discourse analysis—it was examined how discussions might indirectly suggest that AI has the potential for consequences comparable to nuclear disasters, subtly hinting at the possibility of an apocalypse. Then again, these were counted as valid mentions for different categories in the analysis.

Gee's (2014, pp. 80–82) concept of situated meanings provides a valuable framework for understanding how media articles convey significant implications, such as existential threats posed by AI, through indirect references. When articles compare AI risks to nuclear disasters or allude to an apocalypse, they offer contextual cues that trigger readers to construct meanings based on their prior experiences and shared cultural knowledge, in this case, for instance, our historical understanding of nuclear bombs. Such indirect suggestions, though not explicit, evoke associations with large-scale devastation and existential risks. This mutual construction of meaning between the text and the reader justifies counting these indirect references as valid mentions of dangers.

Despite focusing on a limited sample size, this work was characterized by careful, individual attention both in the selection of the articles for investigation, as well as in the processing of the chosen journalistic pieces, as opposed to efficiency-driven, shallow automated analyses. The approach employed is believed to provide substantial added value to the discourse on AI risks.

Taxonomy and classification of AI risks in the literature. In the following, a review of existing taxonomies and frameworks for categorizing AI threats is provided, offering a concise overview of scholarly approaches to classification. However, unlike these systems, which mainly aim to identify all potential dangers systematically, this work takes a different approach. Following the principles of thematic and discourse analysis as described above, the categorization was developed directly from the narratives emerging in articles, ensuring that the framework created reflects the public discourse on AI risks as it appears in the online media.

Numerous studies have presented a comprehensive categorization of AI risks, employing diverse approaches, either with the explicit goal of developing a classification system or as an integral part of the methodology to address the relevant concerns. While not all referenced articles prioritized the question of menaces, they did include them in their discussions in one form or another.

Some authors focused on identifying prevalent topics or specifically types of hazards as outlined in AI guidelines (Jia and Zhang 2022; appliedAI Institute for Europe 2023) or as discussed in academic literature (Clarke and Whittlestone 2022; Hagendorff 2024), while others initiated from theoretical frameworks aimed at dissecting and addressing the complex perils AI presents (Yampolskiy 2016; Tegmark 2018; Cheatham et al. 2019; Russell and Norvig 2020, pp. 1037–57; Schopmans 2022; Ambartsoumean and Roman 2023; Kilian et al. 2023; Federspiel et al. 2023; Lin 2024).

Additionally, efforts were made to provide solutions to these identified threats, ranging from more abstract perspectives to practical guidelines (Turchin et al. 2019; Bécue et al. 2021; Bommasani et al. 2021 Kaminski 2022; Hendrycks and Mazeika 2022; Weidinger et al. 2022; Hendrycks et al. 2023; Shelby et al. 2023; Crabtree et al. 2024) not to mention the categorization outlined in the EU AI Act itself (European Commission 2021). Connected to the topic of regulation, applying the viewpoint of industry and government, Zeng et al. (2024) provide a unified taxonomy rooted in government regulation and company policies.

Beyond explicit undertakings to categorize AI risks and deliver taxonomies, some works provide a form of clustering indirectly by addressing concerns about AI's impact and risks from specific aspects, like in a broader societal or more narrow business operation context, like Acemoglu (2021) and Sharma (2024), respectively.

A select number of existing assessment frameworks addressing AI risks are presented in the following to provide a basis for comparison with the framework developed in this study. It must be emphasized again that the classification proposed in the subsequent section was created following thematic analysis, integrating elements from grounded theory—the categories did not emerge from a literature review but from the analysis of articles and excerpts, representing the narratives *in the way they appear in online media articles*.

To promote diversity and enable meaningful comparison, the first categorization system explored below will bear little resemblance to the one proposed in this paper, offering a

contrasting perspective, while the others will exhibit varying degrees of similarity, allowing for a more reflective analysis across absolutely differing and in many respects aligned frameworks.

Crabtree, McGarry, and Urquhart (2024) classify AI risks by separating them into four domains and/or systemic levels of interaction where they must be addressed: risks in the innovation environment (e.g., novelty, limited understanding of developers, regulatory ambiguity), the internal operating environment, that is, the computational system (e.g., model boundary overreach, system integration issues), the external operating environment, that is, the human system (e.g., user-induced flaws, demands prioritizing automation over accuracy leading to malfunctions), and the regulatory environment (e.g., frequent changes, expertise gaps, compliance impact on development efficiency). This framework reflects the diverse sources and stakeholders of risks throughout AI development, deployment, and regulation, and also presents how "iterable epistopics" provide practical insights into risk management.

Similar to this paper, Thomas et al. (2024) primarily focus on identifying and addressing the "hidden" harms associated with AI while additionally also creating a classification. Nevertheless, while the work presented here centers on the media risk narratives, Thomas et al. explore dangers and detect neglected harms in mainstream AI risk frameworks, including environmental harms, which they highlight as particularly overlooked in risk assessment approaches. Thomas et al. map AI risks across two intersecting dimensions: the scale of harm (individual, collective, societal) and AI supply chain stages (resource extraction, resource processing, deployment). The prior aspect —even if adopting different terminology—addresses the same dimension, the magnitude of potential impact, aligning with the framework proposed in the following section of this paper.

Finally, the most comprehensive framework of AI risks is proposed by Slattery et al. (2024). Their AI Risk Repository consolidates 777 risks derived from 43 taxonomies, organized into a publicly accessible and modifiable database. Built through systematic reviews and expert consultations, the repository employs a best-fit framework synthesis to classify risks into two main taxonomies: a Causal Taxonomy, which categorizes risks by the entity (human or AI), intentionality (intentional or unintentional), and timing (pre- or post-deployment) and a Domain Taxonomy, which organizes risks into seven domains—Discrimination & toxicity, Privacy & security, Misinformation, Malicious actors & misuse, Human-computer interaction, Socioeconomic & environmental, and AI system safety, failures, & limitations—further divided into 23 subdomains. Additionally, while the framework offers a robust classification system, it does not explicitly include the impact, magnitude of potential harms, or probability of risks as formal dimensions. Nonetheless, the authors acknowledge the significance of these factors, particularly for policymaking, and suggest that their integration could be a valuable direction for the future development of their system.

In contrast to the above-enumerated approaches, the focus of this study shifts towards understanding AI risks through the lens of public opinion in several aspects, similar to the work presented by Chuan et al. (2019), Nguyen and Hekman (2022), Nguyen (2023), and Xian et al. (2024). Yet, my investigation concentrates more heavily on dangers and spans a subsequent and exceptionally intense phase in AI development, namely the 'post-ChatGPT era,' characterized by widespread adoption and application of LLMs.

Categorization of AI risks

This section, building on the analysis presented in the previous segment, will introduce the present-day viewpoints from the general public regarding the dangers associated with AI, as reflected in media portrayals.

Given that a significant proportion of the AI risks represented in the online articles are deeply interconnected, the goal was to disentangle them as much as possible according to the narratives in which they are typically discussed in media discourse. Another aim was to incorporate them into a multilevel hierarchical structure based on the various angles and layers presented in the articles.

The *core themes*, representing the first and top level of the hierarchy and denoted by Roman numerals, along with their corresponding subgroups, will be explored in the following subsections. Due to the diverse spectrum of societal risks, the fourth core theme was broken down into four *focus areas*, marked by capital Latin letters—this second hierarchical level, therefore, is conditional. At the smallest increments, on the third and lowest level of the hierarchy, the *specific risk categories* are identified using Arabic numerals. It must also be noted that the expression *'core theme'* does not fit the fifth group on the first hierarchical level, V. Undervalued risks, since that did not emerge 'organically' as a theme addressed in the articles but was 'artificially' constructed in order to bring attention to certain un- or underdiscussed dangers of AI.

Naturally, the classification proposed here is not the only possible one. However, the iterative process, involving repeated reviews of the excerpts in their original article contexts, confirmed the grouping presented in the following, as visualized in Fig. 2.

The online media narratives can be separated into two perspectives, which are not mutually exclusive, namely, concentrating on the causes or the consequences, suggesting an inherent, though unintentional, alignment with mainstream moral philosophy by the authors (or the categorizer, or both). On the one hand, in terms of the origin, the authors' accounts implicitly differentiate based on human intentionality (not considering AI agency), that is, between intentional and unintentional harm. On the other hand, regarding the outcomes, the dimension of their severity is suggested, however, only indistinctly: ranging from limited-, over societal- to existential scale. Altogether, the focus in the case of each excerpt was either rather clearly on the roots or the results, allowing categorization into I. Dependence risks or II. Malicious use risks for the former, and into III. Societal risks or IV. Existential risks for the latter. For clarity, it must be underlined that no implication of any hierarchy between the two perspectives is intended—the core themes merely reflect the authors' mode of portrayal, which was relatively distinct. When it came to risks suggesting limited-scale impact, the question primarily revolved around the intentionality dimension.

Only the specific categories assigned to V. Undervalued risks were isolated from the narrative-based framework to shed additional light on them. However, no suggestion will be made regarding where these should be integrated into the structure given by the four core themes, as they do not constitute a normative but a purely descriptive classification aiming to capture *how these risks are portrayed in online media*. In contrast, the fifth top-level group, V. Undervalued risks, represents a strongly normative framework element, aiming to highlight the problematic nature of the infrequent discussion of its underlying topics in public discourse. Then again, while most of these risks could be thematically classified within the framework established by the earlier core themes, this will not be done, as their most defining feature is their very restricted presentation, both relative to their importance and in absolute terms.

The detailed categorization of AI threats, derived from the online media analysis, is outlined in Table 1.

In addition to identifying whether a topic was simply mentioned (*'mere mentions'*), it was also considered whether a specific topic was discussed in detail (*'deeper coverage'*) within the articles. The distinction between these two was not merely based on the

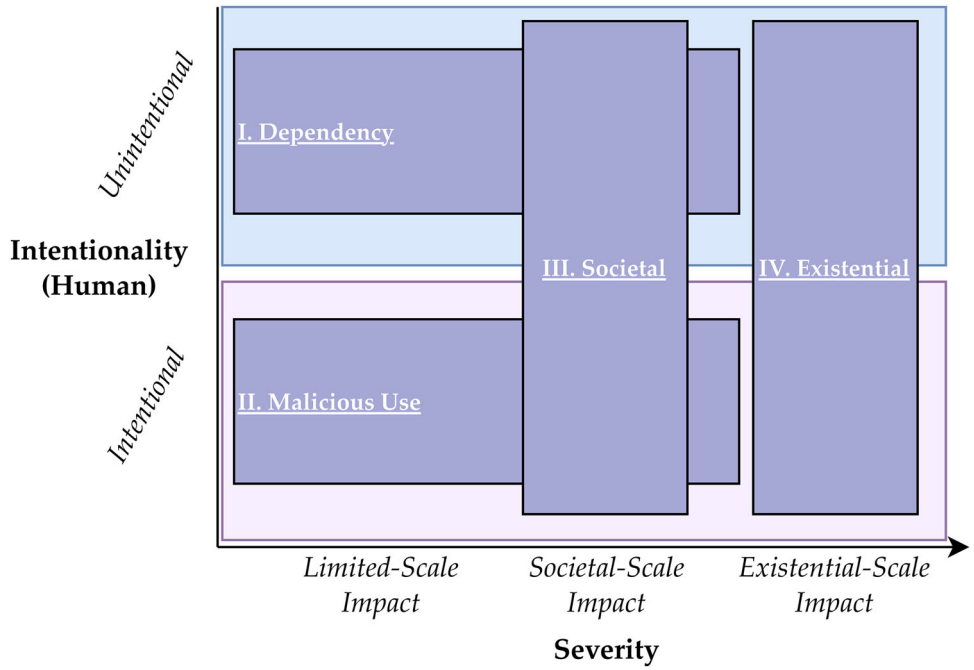

Fig. 2 Key dimensions of online AI risk narratives.

Table 1 AI risk categories identified through the online media review.

I. Dependence		I/1. Malfunction and underperformance
		I/2. Human error
		I/3. Transparency and explainability
		I/4. Control loss and alignment
		I/5. Rogue AI
II. Malicious use		II/1. Cybersecurity
		II/2. Data protection and privacy
		II/3. Autonomous and AI-enabled weaponry
III. Societal	III/A. Sociocultural	III/A/1. Bias, discrimination and inequality
		III/A/2. Information integrity and public trust
		III/A/3. Societal and cultural transformation
		III/A/4. Education and skill degradation
	III/B. Socioeconomic	III/B/1. Workforce displacement
		III/B/2. Economic disruption and market volatility
		III/B/3. Organizational and reputational
		III/B/4. Outsourcing and worker exploitation
	III/C. Sociopolitical and strategic	III/C/1. Political manipulation
		III/C/2. Authoritarianism and privacy erosion
		III/C/3. AI Arms race and cyber warfare
		III/C/4. Power concentration
	III/D. Ethical* and legal	III/D/1. Regulation and oversight
		III/D/2. Accountability and liability
		III/D/3. Intellectual property and copyright
		III/D/4. Negligent and immoral corporate practices
IV. Existential		
V. Undervalued		V/1. Psychology and mental health
		V/2. Environment and sustainability
		V/3. AI Sentience
		+ V/4. Animal Welfare

number of excerpts corresponding to a given category, but also the average depth of the discussion of the specific risk across the articles was considered, therefore in some cases, an article with a few excerpts related to a given risk class was recognized as providing a '*deeper coverage*' on the matter.

While Appendix A provides a statistical overview of the occurrence of all the risk types in the articles, this section—for clarity—details only the quantitatively coverage of V. Undervalued Risks, as these underappreciated dangers are the key concern of this study.

Moreover, in Appendix B, a compilation of the most representative terms ('*Key terms connected*') for each separated class, referenced from the reviewed articles with minimal modification for a formal and standardized presentation, is also provided. Following the approach used for occurrences, this section outlines only the related terms for V. Undervalued Risks.

It must be admitted that there are dangers that still might fit into multiple categories, and the complexity is reflected in the fact that, as previously mentioned, the same terms can refer to various forms of AI risks. Consequently, identical terms may reoccur as keywords for different categories of hazards.

Dependence risks. The core theme I. Dependence Risks illuminates the concerns due to reliance on technology, without suggesting any malevolent intent by the developers or the operators of AI, or the AI entity itself. They presume that dependency might lead to defects or unfulfilled expectations stemming from human or model flaws, often marked by severe opacity. In more severe cases, human command might be diminished and conflicts with human interests might arise, potentially resulting in runaway operations by the artificial agent.

This core theme, however, does not address extreme cases threatening humanity's future: in these instances, consequences of that severity were not part of the authors' narratives. Altogether, in the portrayals of risks being categorized here, online media writers emphasize that the implications stem from technological dependency, shifting focus away from the scale of the resulting events, even if these could be considered societal or existential in scope. It is crucial to underline that this core theme captures accidental outcomes without any harmful intent by humans. To put it another way, the emphasis here lies on the origin, which is, in this case, unintentional, rather than the outcomes.

Dependence—general: Excerpts that could not be classified into the five underlying specific categories fall into this group. These address dependence on AI tools in overarching terms without specific details, implying adverse consequences.

Malfunction and underperformance: it covers technical shortcomings and AI falling short of meeting expectations regarding its performance.

Human error: it emphasizes the aspect of human faults in AI employment or, secondarily, model development, leading to harmful outcomes.

Transparency and explainability: it highlights hazards arising from the difficulty of understanding AI systems and interpreting their outputs.

Control loss and alignment: it addresses concerns about AI operating beyond human supervision and command, potentially conflicting with human values, and evolving in ways that pose significant risks without detailing specific adversarial outcomes.

Rogue AI: it goes beyond a mere loss of control and spans over the severe dangers of AI exceeding human authority, exhibiting manipulative or adversarial behaviors, explicitly acting against human values, and evolving unpredictably, potentially causing widespread and critical disruptions. Extreme cases where this loss of control reaches a level with the potential to fundamentally and destructively transform civilization are already classified as IV. Existential risks.

With regard to these hazards, the idea of AI as a conscious agent arises, though not in a context that raises questions about its potential to suffer. That aspect of risks will be addressed under the category V/3. AI sentience.

Malicious use risks. The core theme II. Malicious Use Risks comprises the results of humans' intentional misuse of AI technologies for harmful purposes. It includes criminal activities, the creation of damaging content, exploitation by hostile actors, and the potential weaponization of AI or enabling weapon creation through AI.

This core theme does not deal with drastically severe scenarios that threaten the trajectory of societal development or the survival of humanity as a whole, as neither humanity-endangering outcomes nor societal-scale risks were implied in the authors' narratives within the excerpts falling under this theme. In other words, the primary focus in these cases is on the cause—which is, in this case, intentional—rather than the effects.

Malicious use—general: Excerpts not fitting into the five specific underlying categories are included in this class. These discuss the malicious use of AI tools broadly, suggesting negative outcomes, without delving into the particulars.

Cybersecurity: it captures perils stemming from the malignant exploitation of AI to target digital systems and infrastructures, creating vulnerabilities that compromise the safety and well-being of individuals without causing direct physical harm.

Data protection and privacy: it covers menaces from unauthorized and malevolent access, use, or exposure of sensitive information, leading to the erosion of personal confidentiality and potential misuse of data. It pertains to non-physical consequences and excludes specific societal-scale intrusions to privacy, as well as copyright or intellectual property-related narratives.

Autonomous and AI-enabled weaponry: it explores the dangers of AI-powered technologies being used by hostile actors in weaponized autonomous applications or facilitating the creation and use of harmful and destructive systems. This class focuses on nefarious exploitations of AI technology causing real physical harm rather than digital impact, specifically by non-state forces, excluding military applications.

Societal risks. The core theme III. Societal Risks spans the broad spectrum of perils that arise due to AI technology on a societal scale. The emphasis in this theme shifts to the consequences rather than the cause. Nevertheless, the impacts discussed here, as described in media narratives, do not reach an existential scale.

Societal—general: Excerpts that do not align with the four focus areas or the associated specific categories are grouped into this class. These discuss the malicious use of AI tools broadly, suggesting negative outcomes without delving into specifics.

Sociocultural risks. It incorporates concerns related to fairness, trust, the transformation of values, education, and the reduction of human capabilities in a rapidly evolving social reality led by AI advancements.

Bias, discrimination, and inequality: it covers a wide range of risks of AI systems reinforcing existing disparities and creating unfair outcomes across various societal and institutional contexts.

Information integrity and public trust: it highlights the dangers of AI contributing to the creation and dissemination of misleading content, with significant societal consequences such as reduced societal trust and the degradation of information integrity.

Societal and cultural transformation: it addresses the widespread effects of AI on the societal status quo and stability, including shifts in cultural norms and social structures, underlining both the risks to societal cohesion and democratic integrity.

Education and skill degradation: it explores the adverse impact of AI on knowledge acquisition, cognitive and general human development, and the thriving of imagination and ingenuity, focusing on concerns about diminished abilities and the potential deterioration of scholarly practices and educational frameworks.

Socioeconomic risks. It comprises the broad economy-related challenges introduced by AI to society, touching on threats arising from shifts in labor dynamics, market stability, organizational integrity, and the exploitation of workers driven by profitability considerations.

Workforce displacement: it captures the impact of AI and automation on labor force activities, including reduced demand for certain roles, transformations in the job market, and potentially leading to vast-scale unemployment.

Economic disruption and market volatility: it sheds light on the diverse economic risks associated with AI, including challenges to market stability, industry transformations, and the unpredictability of AI-driven decision-making systems in financial operations.

Organizational and reputational: it focuses on the risks businesses encounter with AI adoption, including utilization dilemmas, operational risks, financial impacts, and potential harm to public trust and organizational integrity.

Outsourcing and worker exploitation: it reflects the perilous outcomes surrounding the reliance on undervalued and often outsourced labor, where workers facilitating AI systems often endure poor conditions, inadequate compensation, and insufficient occupational norms.

Sociopolitical and strategic risks. It outlines the pitfalls posed by AI to civil liberties, democratic processes, global security, and power dynamics, showcasing concerns over its misuse in governance, information control, military applications, and the concentration of technological authority.

Political manipulation: it includes the complex set of threats posed by AI in distorting information, influencing public perception, and challenging the fairness and integrity of political systems and discourse.

Authoritarianism and privacy erosion: it embraces risks connected to AI-based monitoring systems and the menaces they pose to personal autonomy and freedom, focusing on the misuse by authorities or governments.

AI arms race and cyber warfare: it explores the concerns linked to AI in military and defense systems, highlighting the strategic and security hazards it presents to global stability in both conventional and digital conflicts.

Power concentration: it encompasses the threats associated with the intense accumulation or monopolization of AI control and influence, leading to imbalances in technological authority and resulting in disproportionate power structures.

Ethical* and legal risks. It examines the risks associated with legal, ethical, and corporate governance issues in AI, drawing attention to problems connected to oversight gaps, slow regulation, lack of accountability mechanisms, violation of intellectual property rights, and irresponsible industry conduct.

The asterisk is used to emphasize that the term 'ethics' is applied in its conventional sense—primarily considering human interests while neglecting the well-being of animals, the condition of the natural environment, and the potential experiences of artificial entities.

Ethical* and legal—general: excerpts that fall outside the scope of the four underlying specific categories are assigned to this class, exploring the risks associated with navigating the moral and legal challenges posed by the use of AI technologies, without delving into the nuances.

Regulation and oversight: it explores the challenges related to the governance of AI, focusing on the perils posed by insufficient supervision of development and employment, the rapid pace of technological progress, and the potential for harm due to inadequate regulatory frameworks.

Accountability and liability: it points out the challenges arising from the difficulties of assigning responsibility for the outcomes of AI systems both morally and legally.

Intellectual property and copyright: it captures the ethical and legal challenges surrounding the use of creative works in AI development as outputs, leading to disputes over ownership, consent, and the potential for negative implications for artists, writers, and the broader creative industries.

Negligent and immoral corporate practices: it addresses the risks associated with unregulated and immoral AI use by profit-oriented entities, including ethical lapses and increasing profit through algorithms that present danger to the well-being of individuals and the stability of communities.

Existential risks. The core theme IV. Existential risks, encompasses the most severe risks AI poses to humanity, including potential civilization collapse, human subjugation, and extinction of the entire human race.

Essentially, these dangers can be considered the most radical culmination of either the rogue AI scenarios or those of the malicious, direct or indirect, weaponization of AI. They address the most drastic outcomes, surpassing consequences that could merely be described as societal-scale, since in these, the very survival of society is put into question. Moreover, given the high frequency of mentions and discussions surrounding these perils, they warrant a distinct subset, that is, a separate core theme within this framework.

In many cases, the risk of extinction is not explicitly mentioned in the publications, but a parallel is drawn between the most severe consequences of nuclear and AI technology, clearly conveying the fear of AI being capable of causing annihilation.

There are no specific subcategories to this set of perils, as it is usually not possible to narrow down the narrative to one specific aspect (such as marginalization or destruction of humanity) within an article.

Undervalued risks. To create core theme V. Undervalued risks, as quantitatively outlined in Table 2, the three underlying specific categories were removed from the classification structure presented above. This was done despite the fact that *Psychological and Mental Health*, as well as *Environmental and Sustainability* threats, could have been sorted into the core theme of III. Societal risks.

Table 2 Summary of the media representation of undervalued risks.

		Proportion of ...		
		Articles with mere mentions	**Articles with deeper coverage**	**Total articles with mentions or coverage**
V. Undervalued	V/1. Psychology and mental health	18%	11%	29%
	V/2. Environment and sustainability	9%	9%	18%
	V/3. AI sentience	2%	2%	4%
	+V/4. Animal welfare	0%	0%	0%

Moreover, an additional group, namely V/4. Animal Welfare was introduced, despite no threats posed to animals being even just mentioned in any of the articles reviewed. It was included, nevertheless, because it is firmly asserted here that animals should be a part of our moral framework and the circle of entities whose interests deserve consideration, as also argued by other researchers on AI risks (Ziesche 2021; Bossert and Hagendorff 2021; Hagendorff et al. 2023; Singer and Tse 2023; Bossert and Hagendorff 2023; Coghlan and Parker 2024; Ghose et al. 2024).

As was already indicated, although the three groups—discussed to some degree in the media—could have been categorized within the framework of earlier core themes, they remain excluded and constitute a fifth core theme due to their most defining characteristic: their extremely limited representation, both in relation to their significance and also in absolute terms.

Psychology and mental health. Overall, this category covers the perils posed by AI systems in social, emotional, and mental health contexts, including the potential for emotional dependency and manipulation, exposure to deceptive content, as well as the weakening of real-life connections leading to social isolation, with profound effects on individual well-being and societal stability.

The narratives presented in the media highlight a variety of perils that AI can cause to the human psyche. However, these issues appear in only a limited number of publications—fewer than one-third of the articles in the investigated sample—which contrasts with the significance of mental health-related concerns. What is more, the authors usually just mention them without delving into a deeper evaluation. In the following, a brief overview of the psychological threats discussed in the online publications will be provided.

As AI replaces jobs, many people may face psychological distress, and if their vocation provided them with purpose and fulfillment, it could even lead to a loss of identity (Regalbuto et al 2023; Nolan 2023). Additionally, those outsourced to moderate explicit material for AI training often suffer severe trauma (O'Neil 2023).

AI systems carry considerable dangers to social, emotional, and psychological well-being, particularly in delicate contexts. Generative AI chatbots might promote emotional dependence, reducing empathy and creating unhealthy attachments (Metz 2023; Thomas 2023; Kundu 2024). The features of AI systems that mimic human behavior encourage overtrust and emotional dependency (El Atillah 2024), cultivating unhealthy expectations in relationships (Jones 2024). This reliance has the potential to weaken genuine human connections, leading to detachment and isolation, coupled with a decline in social competence (Marr 2023; Bremmer 2023; Rushkoff 2022).

In fields that are potentially even more vulnerable, such as mental health practice, AI's lack of human sensitivity risks harmful outcomes (Ryan-Mosley 2023), with extreme cases of interactions with chatbots already being reported to lead to suicides (Hale 2023; Sodha 2023). Recommendation algorithms tend to amplify extreme content (Ryan-Mosley 2023), addictively captivating users (El Atillah 2024), often intensifying harmful behaviors and thoughts (Sodha 2023), further impacting the human psyche.

Surprisingly, only a marginal fraction of the articles (in the examined sample, solely one (Gow 2023)) explicitly tackle the issue of AI algorithms on social media platforms, which are employed in manipulating users and exerting a significant impact on their mental health. It is highly likely that this is due to the general notion that as people become accustomed to a technology that incorporates AI, they often no longer perceive it as AI-driven. Accordingly, when authors discuss the dangers, their focus is perhaps not on existing technologies but on future developments.

Key terms connected: *reduced human empathy, decline in human connection, chatbots lead to unhealthy expectations, deep emotional attachment to AI, emotional dependence on AI, emotionally compelling content, addictive algorithms, psychological manipulation, AI in mental health therapy, chatbot-linked suicide, creation of a false sense of importance.*

Environment and sustainability. To sum up, this class discusses the significant ecological risks associated with AI systems, ranging from the extensive use of natural resources and greenhouse gas emissions due to the high energy demand of training and operation, coupled with the water consumption for cooling servers running AI software, to the consequences of scaling up the technology, which could compromise the sustainability and environmental stability.

In media portrayals, the risks associated with ecological and long-term viability issues center around a handful of primary concerns. These are covered in only a small fraction of publications, with fewer than 20% of the articles in the examined sample addressing them, which is disproportionate to the scale and urgency of the environmental challenges. The following is a brief summary of the environmental hazards covered in online articles.

AI systems impose significant environmental costs, requiring large quantities of energy and natural resources while also leading to substantial carbon emissions (Mittelstadt and Wachter 2023). Training powerful models requires massive server farms, leading to high electricity consumption (Hunt, 2023; McCallum 2023) and a significant carbon footprint due to the energy-intensive nature of these computations (Baxter and Schlesinger 2023). Additionally, cooling systems for these models use vast amounts of water, exacerbating the depletion of water sources (Caballar 2024), especially in vulnerable regions (Isik et al. 2024). The development of LLMs and facilitating high-performance computing (Ryan-Mosley 2023; Barrett and Hendrix 2023) also rely heavily on rare earth metals, further straining global resources (Rushkoff 2022). The ongoing trend of training increasingly larger models only amplifies these environmental challenges, placing additional pressure on the planet's sustainability (Wai 2024).

Key terms connected: *environmental impact, carbon footprint, energy consumption, electricity usage, energy-intensive platforms, carbon emissions, water use, fast depletion of water sources from vulnerable parts of the planet, massive server farms, resource-intensive datasets/models, the trend to train bigger models, consume a huge volume of hardware/natural resources, computers and servers require massive amounts of rare earth metals, a disservice to the planet.*

AI sentience. In essence, this group addresses the dilemmas posed by the hypothetical potential for AI to develop subjective, phenomenological experiences and the capability to suffer, raising questions about their moral status.

In the overwhelming majority of publications on the risks of this technology, AI sentience-related menaces remain ignored. These concerns appear in less than 5% of the articles in the analyzed sample. The following offers a compilation of the threats related to AI sentience in online media.

As AI technology advances, the possibility of artificial agents achieving sentience—experiencing emotions and sensations—becomes a more and more realistic scenario. Determining whether AI deserves ethical consideration, similar to the separate moral statuses of humans and animals, will present a significant challenge. The risk lies in the potential mistreatment of sentient AI, either unintentionally or intentionally, if proper rights are not granted (El Atillah 2024).

Even though *Rogue AI* risks and the related scenarios might indicate a level of self-awareness, the focus in that category was

entirely on the impact the agent exerts on humans, not on the implications of possessing consciousness for the entity itself.

Key terms connected: *AI becomes sentient, AI achieves sentience, humans mistreat sentient AI, moral considerations of sentient AI, AI systems reaching the level of sentience/consciousness/self-awareness.*

Animal welfare. It is not mentioned or suggested in any of the reviewed articles, even in the subtlest manner, that the interests of animals should be taken into consideration when examining the potential dangers of AI technology.

Key terms connected: none.

Evaluation of the media review, categorization, and underappreciated risks

This section evaluates the proposed AI threat categorization and the representation of the categories in online media articles, as these portrayals, in turn, reflect public opinion. The discussion begins by relating the developed classification system to other established frameworks before turning to the key concern of a detailed examination of the neglected AI risks—an examination that forms the primary objective of this paper. While potential explanations for the disregard of these topics will be explored, an indicative collection of potential risks, rather than an exhaustive list, will also be outlined.

Reviewing online news articles is an established method for investigating overall media discourse, as demonstrated in the work of other researchers (Chuan et al. 2019; Sun et al. 2020; Nguyen and Hekman 2022; Nguyen 2023; Xian et al. 2024). While relying solely on these sources does not offer a comprehensive analysis of the media discourse on AI risks due to the exclusion of various other platforms, online media articles sufficiently illustrate the public dialogue about this topic. They also effectively underscore that the dominant narrative is predominantly centered on human concerns, providing insight into the media's approach.

The discourse fails to address the role of animals and the environment, not even acknowledging their positions within feedback loops that could impact human civilization. Likewise, the matter of sentient AI has barely been explored. Clearly, while specific reports may cover these topics, their absence from general-interest articles may result in the general public remaining uninformed about these pitfalls. This lack of awareness persists unless individuals seek out this information intentionally due to a personal interest in these matters.

Then again, this anthropocentric approach is hardly unexpected (Owe and Baum 2021; Hagendorff 2022; Rigley et al. 2023, pp. 844–848), however, it still misses the conditions of the overwhelming majority of the animals currently living on the planet on the one hand, and our environment as a system on the other. To say nothing of the moral catastrophe that might be hypothetically caused due to the emergence of sentience or consciousness in machines.

Comparison of the proposed categorization with existing frameworks. Building on the overview of AI risk taxonomies in the section "Taxonomy and classification of AI risks in the literature", one parallel deserves further attention.

From the structure developed by Slattery et al. (2024), many comparisons can be made with the classification framework introduced in this paper. Their theoretical framework operates in the three-dimensional matrix of entity, intentionality, and timing. The categorization derived from the media analysis also highlighted the relevance of the causal dimension of intentionality, but the core themes developed, as illustrated in Fig. 2, did not exhaust all logical possibilities of that theoretical matrix, as my empirical approach prioritized the capturing of the narrative structures articulated by the authors of the articles, and these perspectives were not prominently represented in the textual corpus.

The entity dimension appeared only marginally, however, the introduction of a temporal perspective—distinguishing between pre- and post-deployment—could have enriched the analysis. Having said that, drawing on the terminology of Slattery et al., this study focused on the Domain Taxonomy rather than the Causal Taxonomy. Therefore, adding another dimension to the latter would have unnecessarily complicated the categorization into core themes.

Furthermore, Slattery et al. emphasize the importance of incorporating the magnitude or scale of impact, supporting the two-dimensional framing of media discourse based on intentionality and severity, as suggested in this paper. While this does not confirm that these dimensions represented the dominant narrative threads, it does demonstrate that their separation along these lines provides a coherent and meaningful framework for analysis.

About the anthropocentricity. The representation of different types of dangers in online media articles clearly shows that the ongoing discussion regarding the technology's potential dangers is overwhelmingly human-centered. This is clearly reflected in the fact that only 5% of the reviewed publications mention any entity other than humans that might be capable of suffering, namely those that ponder upon the hypothetical possibility of AI gaining sentience and/or consciousness. What is more, only one article in the entire sample explicitly regards artificial agents as moral subjects, whereas the interests of animals are ignored altogether. Even though *Rogue AI* risks might suggest a degree of self-awareness, these narratives make no reference whatsoever to the potential of these entities to endure suffering.

Anthropocentricity is also evidenced by the way nearly one-fifth of the reviewed publications, which address the technology's environmental dangers, frame this issue. While ecological and sustainability concerns might also stem from non-anthropocentric, such as planet-focused roots, these articles provide no indication—and thus offer no basis for believing—that the authors consider the natural world as an independent moral subject. In other terms, the focus of environmental issues seems to lie in their human-centered impacts.

Disregard for animals. What stands out about the previously revealed human-centeredness is that none of these articles mention the harm that the technology may pose to non-human animals by any means. Even the media pieces that address environmental issues do not mention wild animals or refer to any species other than Homo sapiens in any way.

Potential explanations for the neglect of animals. This observation, which is the absolute neglect of AI's potential impact on animal welfare, resonates perfectly with what was stated previously regarding ethics (with an asterisk). Namely, moral considerations, as a rule, refer solely to issues in which harm is potentially being done against human beings, either directly or indirectly. This investigation has provided evidence that our moral sentiments and the prevailing methods of risk assessment in our civilization are profoundly deficient. From a perspective that aims to consider the interests of all beings capable of experiencing physical and mental agony, our decision-making processes fall significantly short.

In Singerian terms, most people in our civilization maintain a speciesist bias (Singer and Tse 2023, p. 1–5), which is obviously

reflected in the examined articles. Even if the term speciesist bias appears in the unfortunately rather marginal segment of AI ethics that encounters animal welfare issues (Ziesche 2021; Bossert and Hagendorff 2021; Hagendorff et al. 2023; Singer and Tse 2023; Bossert and Hagendorff 2023; Coghlan and Parker 2024; Ghose et al. 2024) and the overwhelming majority of the reviewed news coverage addresses the difficulties rooted in AI-amplified bias, it is abundantly clear from the context that the authors of these articles were using the phrase in a human-centric manner.

The same conclusion is also apparent from the nature of media reports, where being capable of raising a novel aspect that has not yet been covered or is not regularly covered is a driving force. In spite of this, no AI concerns regarding animals were mentioned, they linked to bias by any other means whatsoever.

Be that as it may, it must be mentioned that a large part of the bias-related realizations in the examined articles applies not just to human minorities and other societal groups but also to animals being exploited in intensive animal farming. In essence, the manner in which we engage in discussions and think about agricultural animals regularly reflects an exploitative perspective that will be fed into the systems, the bias inherent in the input data will propagate to the AI systems similarly to racism and sexism (Hagendorff et al. 2023; Singer and Tse 2023, pp. 546–547; Ghose et al. 2024).

Apart from the prevalent speciesist bias in society leading to the consideration of non-humans being sidelined, further ideas might help to explain the disregard of animal interests in public opinion, though they remain speculative and without direct literature support. Regarding wild animals, limited awareness of AI's indirect effects on ecosystems and the complexity of tracing its ecological impact may contribute to this neglect. Economic interests at both private and governmental levels, such as AI-driven factory farming, prioritize efficiency and profit over animal welfare, shaping how these issues are portrayed in the media through advertising revenue or corporate partnerships. Powerful lobbying efforts by industries reliant on AI may further suppress narratives that highlight the risks to animals, promoting favorable public perception and financial gain at the national economic and microeconomic scales. On top of this, the lack of explicit regulatory provisions for AI's impact on non-human life creates a permissive environment for these industrial animal agriculture companies to potentially cause suffering to animals. Additionally, techno-optimism may cultivate a belief that AI will eventually benefit all living beings, even if present risks to animals are overlooked.

Identifying the overlooked animal risks. Although risks posed to animals are not presented in the online media articles, it was shown that some conclusions can still be drawn based on their human-related counterparts. However, AI technology poses further dangers to these sentient beings. Singer and Tse sort these threats along two lines: from one angle, depending on whether the AI that causes harm was designed to interact with animals or not, and from another, based on whether the effect was exerted directly or indirectly by the system (Singer and Tse 2023, p. 541).

Fitting examples of AIs engineered to interact with animals, thereby directly impacting them, include agricultural applications such as handling chickens and milking cows on poultry and dairy farms. Other examples could be AI-driven systems managing various species in zoos, operating as part of a pet training system, or, as a deliberate instance aimed at killing animals: hunting drones. (Singer and Tse 2023, p. 541).

Illustrations of accidental but nevertheless direct AI-animal interactions would be autonomous vehicles that might hit and, as a result, injure or kill these beings. Similarly, housekeeping robots, programmed not to hurt the pet companions of the owner, could still cause damage to other animals they come in contact with (Singer and Tse 2023, p. 541).

Instances of AI systems indirectly affecting animals have already been mentioned, but to name one more, algorithms recommending or (not) restricting animal cruelty videos can impact interest in such content, influencing attitudes toward animals and possibly motivating harmful actions (Singer and Tse 2023, p. 541).

In some cases, it is rather difficult to clearly disentangle risks along the proposed dimensions, especially in the long run. As researchers and developers in the field of Natural Language Processing (NLP) regard speciesist bias as an insignificant issue and do not consider it an ethical challenge to tackle, it remains unaddressed in datasets, and as a result, LLMs have a general tendency to produce harm-inducing outputs about animals as a default stance, even in the absence of hostile prompts. This could not merely strengthen the speciesist attitudes of the interactors and worsen the situation of animals overall as the inferiority of animals will be rooted further into the cultural fabric, but assuming that these models will be integrated into embodied machines that have the potential to interact with the physical world, they will be able to inflict direct damage on animals (Takeshita and Rzepka, 2024, pp. 10–11).

Coghlan and Parker (2023) provide a "Harm Framework for Animals and AI" in which they differentiate human-caused harms to animals based on deliberateness. According to this framework, intentional actions can be socially condemned and/or illegal (e.g., AI-powered drones used to track animals for illegal wildlife trade) or socially accepted and/or legal (e.g., AI-driven agriculture enhances control but causes further animal suffering). Unintentional actions may have direct (e.g., technological failures in AI-enabled farming cause suffering to animals) or indirect (e.g., AI expansion disrupts animals' natural habitats) consequences. (Coghlan and Parker, 2023, 12).

The authors also classify foregone benefits, which category reflects on losses from inaction or missed opportunities. They also discuss unintentional indirect effects, including harms from estrangement—the concern that AI in farming may distance humans from animals, reducing care and increasing stress—and epistemic harms, where AI may reinforce beliefs that animals lack moral value, perpetuating and worsening animals' situation (Coghlan and Parker, 2023, 20–22), as already described above.

Finally, it may be concluded that the interests of animals are, regrettably, completely ignored in the public conversation regarding AI risks, despite the wide-ranging implications the technology holds for these beings. At least, this conclusion can be drawn based on the reviewed sample, which broadly reflects public opinion.

Disregard for the environment and sustainability

Potential explanations for the neglect of the environment. No more than merely one-fifth of the reviewed articles deal with the topic of AI's impact on the environment, what is more, half of these only touch upon the topic. This is peculiar considering the fact that, typically, the issues of sustainability and climate change are prominently displayed in the press (Hase et al. 2021; Schäfer and Painter 2021).

In some cases, the authors of the reviewed pieces consider the "*people* and the planet" (Mittelstadt and Wachter, 2023; Gregory and Kleinman 2023) or the "collapse of the environment which *sustains us*" (Clarke 2023), stressing nature's role in supporting mankind's existence on Earth. Other authors remain superficial by indicating climate costs and carbon emissions in general or talk in broad terms about different kinds of consumptions typically connected to the environmental impact without elucidating who will eventually bear the potential consequences

of excess resource use as an afflicted party (McCallum et al. 2023; Baxter and Schlesinger 2023; Ryan-Mosley 2023).

It is no surprise, bearing in mind the conventional interpretation of the term 'ethics' and the line of thinking it signifies, that there is no indication that would allow one to conclude that any of the authors consider either the environment as a whole itself or any non-human living organism (not even the natural fauna) being the moral subject in this question despite the fact that they will just as well and already do experience any consequences of climatic disruption.

While ecological issues, such as climate change and sustainability, are generally prevalent topics in the media (Hase et al. 2021; Schäfer and Painter 2021), as it was shown, AI's potential harm to the environment remains underrepresented in online articles. As speculative explanations, this can be attributed to limited awareness of AI's energy consumption and ecological impact. Similar to the previously noted disregard for animal interests, the economic priorities of AI industries might distort media narratives and suppress coverage of environmental risks through corporate influence and lobbying. As already argued above, techno-optimism probably contributes to the belief that AI will ultimately solve environmental challenges, overshadowing current risks.

Identifying the overlooked environmental and sustainability risks. Regarding the ecological consequences, the risks associated with the entire AI supply chain encompass various factors (C. Thomas et al. 2024). These include the significant power consumption required for training and operating AI models, which—due to the energy source distribution in electricity production still heavily relying on fossil fuels—leads to vast-scale greenhouse gas emissions. Additionally, there is the technology's considerable water footprint, often associated with the cooling processes required to prevent overheating in data center infrastructure. Finally, the demand for specific raw materials, such as lithium or cobalt, must also be taken into consideration. These and even more challenges arise at various stages throughout the life cycle of an AI solution or service, flagging concerns regarding the technology's impact on the environment and its overall sustainability (Ligozat et al. 2022).

The construction of functional AI systems and powering their computations comes at the cost of resource exploitation, including energy carriers and ore deposits. The resource-intensive activities include large-scale mineral extraction and energy-demanding processes for training models, especially in NLP and computer vision. These practices contribute to significant environmental degradation, such as deforestation, toxic waste, and expanded carbon footprints. Despite these impacts, the tech industry rarely bears the ecological costs of its operations and continues to reshape and exploit the planet's resources. (Crawford 2021, pp. 15–51).

As has been demonstrated, environmental concerns are, even if marginally, acknowledged among AI risks in the online media, yet predominantly from an anthropocentric perspective. The discussion concentrates on the impact on humans, failing to reflect on AI technology's effect on the natural world itself.

Disregard for sentient AI

Potential explanations for the neglect of sentient AI. Despite the dominance of the "machine question" in the philosophical discourse on AI (Harris and Anthis 2021), that is, whether the AI can or cannot attain the ability to have subject experiences and in spite of the fact that artificial consciousness is a continually recurring theme in science fiction works (Alvero and Peña 2023), apparently, the matter is sidelined in the public's perception, as evidenced by its underrepresentation in media coverage. One factor that could explain this phenomenon is that these concerns are implicitly included in the journalists' ruminations relating to existential concerns.

Speculatively, another reason why AI sentience risks are overlooked in the media might be that discussions surrounding AI's potential for consciousness often lack clear definitions and remain highly speculative, which ambiguity makes it difficult for journalists to present these risks in a tangible, relatable manner. Besides, the focus on more immediate, practical concerns—such as AI's impact on jobs, privacy, and security—along with existential risks, which are potentially catastrophic for the reader and thus capture their attention more easily, dominates media narratives, leaving less room for the exploration of AI-sentience-related issues. Grabbing and holding the reader's interest is a key aim for many media outlets, and these dramatic, direct risks, supported by popular culture, are more likely to achieve that goal.

Only one of the reviewed publications handles the topic of AI's potential for gaining sentience and consciousness in depth (El Atillah 2024), however, this article covers all the key consequences of this hypothetical scenario, namely the moral considerations of this development, addressing the possibility of artificial entities being mistreated, also raising the question of granting rights to the machines.

Identifying the overlooked AI sentience risks. However, the implications of sentience go beyond the mere capacity for suffering as we could imagine in everyday terms. The fundamental argument here is that the simulated minds—supposedly possessing human-like phenomenological experiences—would not simply outnumber biological humans, these would constitute the near-total majority of the minds that ever existed. (Bostrom 2003) The 'astronomical numbers' of these simulations could lead to the amount of suffering that would vastly exceed the misery manifested throughout the entire history of our planet, called the 'astronomical suffering outcome' (Sotala and Gloor 2017). This could occur, for instance, to enable an agent (not even necessarily having reached the level of superintelligence) to acquire knowledge about human behavior by conducting wide-ranging psychological and sociological experiments on conscious simulated minds that may theoretically merit moral status. This is only one manifestation of a hypothetical catastrophic scenario that involves suffering within the AI itself—that is, computations that are ethically concerning due to the inherent qualities of these processes themselves, independent of any impact on the external world—commonly referred to as 'mind crime' (Bostrom 2014, p. 152; Sotala and Gloor 2017; Bostrom et al. 2020, p. 15)

Further but less drastic implications of sentient AI could be explored, however, due to the strongly hypothetical nature of these menaces, the sole instance where extraordinary severity compensates for minimal likelihood was elaborated upon here. It is maintained that priority should lie elsewhere, but these perils must not be forgotten nonetheless.

All things considered, apart from one single article in the sample that reports on the potential sensations and subjective experiences that these AI agents might endure, it can be concluded that the entire aspect of the potential emergence of a capability for suffering in artificial entities and the implications that development would give rise to regarding the consideration of their interests, is highly overlooked in the public discussion.

Disregard for human psychology

Potential explanations for the neglect of human psychology. While psychological concerns receive appreciable recognition in the public discourse, compared to those even more neglected topics

related to animals, the environment, or sentient AI, the topic's significance still seems to be underestimated. In spite of the anthropocentric emphasis of the entire risk assessment in the media, this particular aspect that deals with AI's power to severely affect the mental health of people and even to change the human identity as we know it now, receives less attention than its weight would indicate.

The critical subjects emerging in the media narratives span addictive algorithms and manipulation on social media platforms (Gow 2023; Ryan-Mosley 2023), diminished social interactions, and reliance on social AI companions (Bremmer 2023), which are both caused by and contribute to a loss of empathy (M. Thomas, 2023), and the degradation of human connections (Rushkoff 2022; Marr 2023), resulting in growing isolation (El Atillah 2024) and even a detachment from reality (Kundu 2024). In connection with this, concerns related to the false attribution of human-like qualities to AI (El Atillah 2024), resulting in unhealthy expectations in human relationships (Jones 2024), inappropriate emotional dependency on these systems (Metz 2023; Kundu 2024), in some cases leading to the control and radicalization of human individuals (Sodha 2023), also arise. Additionally, the authors address the impacts of job displacement on individual drive and purpose in life (Regalbuto et al. 2023; Nolan 2023), psychological effects on data labelers and content moderators (O'Neil 2023), employment of AI in mental health therapy (Ryan-Mosley 2023), and incidents of chatbot-linked suicide (Hale 2023; Sodha 2023).

Surprisingly, only a marginal fraction of the articles explicitly tackle the issue of AI algorithms on social media platforms, which are employed in manipulating users and exerting a significant impact on their mental health. One possible explanation could be that, as individuals become accustomed to AI-integrated technologies, they tend to cease recognizing them as AI-driven. Consequently, discussions on the potential dangers of AI often focus more on speculative future developments than on existing technologies.

The fact that fewer than a mere third of the reviewed articles mention and roughly 10% of them discuss in any real depth the threats AI poses to the human psyche demonstrates the insufficient representation of these threats in the online media. For one thing, this is far too few, in contrast to how widely recognized it is that this emerging technology is impacting our minds in various ways. For another, the manner in which these considerations are pointed out in the articles, namely, as side remarks and in the form of subtle indications, illuminates that there is no established way of addressing them.

In accordance with this, it is worth noting that the key terms presented in the part of the previous section reflecting on the portrayal of *Psychological and mental health risks* were remarkably more diverse than in the case of any other specific category. Although the number of expressions listed in that segment is not outstanding—in contrast to most other danger groups, which included a considerable number of synonyms—the terms associated with psychological perils reflect exceptional variability.

To put it another way, for the most part, these hazards are linked to other, more prevalent narratives in the reviewed publication, and the different articles cover a broad spectrum of perils that leave the impression of being the products of the authors' creativity, as they are absolutely diverging. This could be a sign of an emerging perspective that, while not yet fully defined, is beginning to capture people's attention because they recognize its importance. It is likely coupled with the fact that, as the highlights from scholarly works on AI risks to human psychology —addressed shortly in the upcoming segment—will demonstrate, the range of dangers is immensely broad.

At first glance, one might assume that the more risks associated with a topic, the more attention it will receive. However, it may be the case that while cybersecurity and employment-related risks—two examples among the more frequently discussed categories—are easier to convey with a few phrases, the breadth of mental health challenges is so extensive that authors often avoid attempting to summarize them concisely in online media publications, maybe not even consciously.

In a speculative manner, psychological risks may be underestimated because they are often less tangible and harder to measure compared to more immediate physical concerns. Media outlets tend to focus on more visible, urgent issues, while the long-term, subtle effects of AI on mental health and cognition are harder to convey and may not capture immediate public attention, despite their potential significance.

Identifying the overlooked AI sentience risks. As extensive as the portrayal of human psyche-related threats in the media might seem, many aspects and nuances remain uncovered in the articles that made up the sample for this investigation. Fiske, Henningsen, and Buyx (2019, p. 6–7) besides reporting on the immediate consequences of AI implementation in a therapy context, also review a broad spectrum of fears regarding the long-term effects of these modes of application, partly overlapping and affirming the perspectives of online media articles, as summarized in the following.

The potential for patients to become overly attached to AI applications, such as robots designed to reduce loneliness or provide emotional comfort, raises concerns about dependency (Cresswell, Cunningham-Burley, and Sheikh 2018). AI systems, particularly 'care bots,' could alter human identity in specific aspects as caretaking responsibilities are increasingly outsourced to machines. Moreover, a similar effect could be exerted by intelligent robots on the self-concept and self-awareness of humans, as they will transform interpersonal relationships, as bonds with machines (sexual partner robots can serve as an illustrative example) will possibly weaken human-to-human connections. (Fiske et al. 2019, p. 6–7).

Changing social expectations and communication practices also shed light on the possible influence of AI technology on social psychology. For instance, users often engage rudely with virtual assistants as they lack emotions and thus will not feel hurt, presumably affecting their attitudes toward other individuals as well. Complications are even more unsettling in the case of the already mentioned sex robots, as the application of these has the potential to reinforce objectification and thus contribute to sexual violence. Having said that, the other end of this spectrum also presents hazards, namely the tendency to attribute human traits to these systems or believe they possess individual identities. (Fiske et al. 2019, p. 6–7).

It is also argued that the implementation of AI in various embodied forms might exacerbate reductionist thinking in the mental health profession by providing surface-level treatments for illnesses, overshadowing the complex, biological-psychological-social understanding of pathological conditions. (Fiske et al. 2019, p. 6–7).

AI systems give rise to several other psychological risks apart from the ones in a professional therapy context. These include inducing negative emotions (such as fear or sadness), exacerbating mental health issues (e.g., depression or anxiety), fostering addiction (such as social media and gaming), and offering harmful or misguided mental health advice (e.g., chatbots as therapists) (Pałka 2023, p. 11–12).

It must be emphasized, though, that the aim of this study is not to provide an exhaustive account of the psychological risks associated with AI technology but to highlight the discrepancy

between the extensive scope and importance of these perils and their relative underrepresentation in online media reports.

Overall, the impact of AI on mental health and its potential to alter human identity is only marginally addressed in media coverage, with few articles touching on the issue, often in a rather brief manner. However, the authors approach them from diverse perspectives, without relying on well-established ways of expression—likely due to their absence—unlike in the case of the majority of other AI risk categories. This suggests a growing awareness of its significance, which, however, still needs to penetrate the mainstream discourse.

Conclusion

In this study, the focal points and the ignored dimensions of AI risks were identified, with a particular emphasis on the latter, including animal welfare, environmental effects, AI sentience, and the technology's psychological impact on humans. The presented review of online news articles revealed predominant attention to human-related concerns in the public discourse about AI, often sidelining broader implications for other entities and the environment.

In this analysis, 56 articles were systematically selected from the period of the 'post-ChatGPT era' and subjected to meticulous manual review, utilizing discourse and thematic analysis—particularly framework analysis and integrating elements from grounded theory—as key methodological approaches to exceed the limitations of simplistic, automated examination. Attention was centered on individually and systematically selecting general-interest online media publications and conducting a thorough, context-sensitive evaluation of excerpts extracted from the articles. This approach facilitated a nuanced analysis and consideration of implicit content and the underlying implications within the textual data, which were coded systematically to iteratively develop a categorization system.

Despite the anthropocentric focus revealed in the review, public discussion still undervalues AI's psychological effects on individuals. However, the limited yet varied coverage of mental and emotional concerns suggests that this topic is gaining attention, indicating an emerging awareness of its significance. The analysis underscores the importance of redistributing emphasis across various human-centric perils associated with AI to develop a more balanced and effective strategy for addressing psychological hazards. This approach can contribute to more proficient prevention and management of AI's adverse impacts.

Through the categorization of concerns revealed in the written online media coverage, the focal points, as well as blind spots in the current discourse on AI dangers, were identified, highlighting the necessity for a more inclusive approach in risk assessment frameworks. This paper argues for the extension of considerations beyond human-centric issues in the public conversation by journalists and public intellectuals to encompass primarily the interests of animals and the stability of the environment and, speculatively, the moral status of artificial agents potentially achieving sentience.

By broadening the ethical scope, it is possible to address the complex challenges AI presents more effectively, ensuring that technological advancements do not come at the cost of harming sentient entities or the ecological systems that sustain life. The conclusions highlight the imperative for further inquiries into the overlooked aspects identified in this article, ensuring a comprehensive understanding and mitigation of AI risks.

Based on the findings of this study, a shift in the conceptualization and discussion of AI risks is advocated. It is argued that primary attention must be directed, both in public discourse and future scholarly research, towards investigating AI's impact on animals and the environment. Additionally, consideration should be given, even though only tangentially, to the hypothetical possibility of the moral inclusion of sentient artificial agents. Within anthropocentric risks, the focus should be adjusted to place greater emphasis on psychological effects on humans.

To conclude, this paper calls for an expanded dialogue on AI risks, one that fully recognizes and addresses the diverse spectrum of threats inherent in the advancement of AI to ensure their more effective mitigation. This includes moving beyond both far-fetched speculative scenarios and the most apparent dangers to embrace a comprehensive evaluation of potential hazards for animals, human psychology, the environment, and, more hypothetically, artificial agents gaining sentience—none of which are adequately addressed in the public discourse, as reflected in the written online media.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Received: 22 May 2024; Accepted: 25 March 2025;
Published online: 23 April 2025

References

Acemoglu D (2021) Harms of AI. National Bureau of Economic Research, Cambridge. https://www.nber.org/papers/w29247

Alvero AJ, Peña C (2023) AI sentience and socioculture. J Soc Comput 4(3):205–220. https://doi.org/10.23919/JSC.2023.0021

Ambartsoumean VM, Roman VY (2023) AI risk skepticism, a comprehensive survey. Preprint at https://doi.org/10.48550/arXiv.2303.03885

Andrews K, Birch J, Sebo J, Sims T (2024) Background to the New York Declaration on animal consciousness. The New York Declaration on Animal Consciousness

AppliedAI Institute for Europe (2023) AI act: risk classification of AI systems from a practical perspective. https://www.appliedai.de/assets/files/AI-Act_WhitePaper_final_CMYK_ENG.pdf

Balcombe J (2017) What a fish knows: the inner lives of our underwater cousins. Oneworld Publications

Barrett PM, Hendrix J (2023) We must address the AI risks right in front of us today. The Hill. https://thehill.com/opinion/technology/4079054-we-must-address-the-ai-risks-right-in-front-of-us-today/

Baxter K, Schlesinger Y (2023) Managing the risks of generative AI. Harvard Business Review. https://hbr.org/2023/06/managing-the-risks-of-generative-ai

Bécue A, Praça I, Gama J (2021) Artificial intelligence, cyber-threats and industry 4.0: challenges and opportunities. Artif Intell Rev 54(5):3849–3886. https://doi.org/10.1007/s10462-020-09942-2

Bommasani R, Drew AH, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS et al. (2021) On the opportunities and risks of foundation models. Preprint at https://doi.org/10.48550/arXiv.2108.07258

Bortolotti, L, M Mameli and, and A Blasimme (2013) "Sentience, moral relevance of." In: The international encyclopedia of ethics. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444367072.wbiee027

Bossert L, Hagendorff T (2021) Animals and AI. The role of animals in AI research and application—an overview and ethical evaluation. Technol Soc 67(November):101678. https://doi.org/10.1016/j.techsoc.2021.101678

Bossert L, Hagendorff T (2023) The ethics of sustainable AI: why animals (should) matter for a sustainable use of AI. Sustain Dev 31(5):3459–3467. https://doi.org/10.1002/sd.2596

Bostrom N (2003) Are we living in a computer simulation? Philos Q 53(211):243–255

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford, New York

Bostrom N, Dafoe A, Flynn C (2020) Public policy and superintelligent AI: a vector field approach. In: Matthew Liao S (ed) Ethics of artificial intelligence. Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0011

Braithwaite V (2010) Do fish feel pain? Do fish feel pain? Oxford University Press, New York

Bremmer I (2023) The four big risks from artificial intelligence. The Philippine Star. https://www.philstar.com/opinion/2023/06/13/2273499/thefourbigrisks-artificial-intelligence

Caballar R (2024) 10 AI dangers and risks and how to manage them. IBM. https://www.ibm.com/blog/10-ai-dangers-and-risks-and-how-to-manage-them/

Ceran B, Kedia N, Corman SR, Davulcu H (2015) Story detection using generalized concepts and relations. In: 2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 942–949. https://doi.org/10.1145/2808797.2809312

Cheatham B, Javanmardian K, Samandari H (2019) Confronting the risks of artificial intelligence. McKinsey Q. 2(38):1–9
Chuan C-H, Tsai W-HS, Cho SY (2019) Framing artificial intelligence in American newspapers. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. AIES'19. Association for Computing Machinery, New York, pp 339–44. https://doi.org/10.1145/3306618.3314285
Clarke S, Whittlestone J (2022) A survey of the potential long-term impacts of AI. In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, pp 192–202. https://doi.org/10.1145/3514094.3534131
Clarke, T. 2023. AI risks are something humanity can manage—let's get working on how it can save us from ourselves. Sky News. https://news.sky.com/story/ai-risks-are-something-humanity-can-manage-lets-get-working-on-how-it-can-save-us-from-ourselves-12901265
Coghlan S, Parker C (2023) Harm to nonhuman animals from AI: a systematic account and framework. Philos Technol 36(2):25. https://doi.org/10.1007/s13347-023-00627-6
Coghlan S, Parker C (2024) Helping and not harming animals with AI. Philos Technol 37(1):20. https://doi.org/10.1007/s13347-024-00712-4
Conway M (2006) The subjective precision of computers: a methodological comparison with human coding in content analysis. J Mass Commun Q 83(1):186–200. https://doi.org/10.1177/107769900608300112
Crabtree A, McGarry G, Urquhart L (2024) AI and the iterable epistopics of risk. AI & SOCIETY. https://doi.org/10.1007/s00146-024-02021-y
Crawford K (2021) The atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press
Cresswell K, Cunningham-Burley S, Sheikh A (2018) Health care robotics: qualitative exploration of key challenges and future directions. J Med Internet Res 20(7):e10410. https://doi.org/10.2196/10410
El Atillah I (2024) 5 of the Most damaging ways AI could harm humanity, according to MIT experts. Euronews. https://www.euronews.com/next/2024/09/01/ai-could-go-wrong-in-700-ways-according-to-mit-experts-these-are-5-of-the-most-harmful-hum
European Commission (2021) EU AI Act. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF
European Union (2007) Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community. European Union, Lisbon. http://publications.europa.eu/resource/cellar/688a7a98-3110-4ffe-a6b3-8972d8445325.0007.01/DOC_19
Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D (2023) Threats by artificial intelligence to human health and human existence. BMJ Glob Health 8(5):e010435. https://doi.org/10.1136/bmjgh-2022-010435
Fiske A, Henningsen P, Buyx A (2019) Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 21(5):e13216. https://doi.org/10.2196/13216
Gee JP (2014) An introduction to discourse analysis: theory and method, 4th ed. Routledge, London. https://doi.org/10.4324/9781315819679
Ghose S, Yip Fai Tse KR, Jeff Sebo PS (2024) The case for animal-friendly AI. Preprint at https://doi.org/10.48550/arXiv.2403.01199
Gibbs GR (2007) Analyzing qualitative data. SAGE Publications, Ltd, https://doi.org/10.4135/9781849208574
Gloor D, Althaus L (2016) Reducing risks of astronomical suffering: a neglected priority. Center on Long-Term Risk (blog). https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/
Gow G (2023) Top 5 AI risks in the era of ChatGPT and generative AI. Forbes. https://www.forbes.com/sites/glenngow/2023/04/09/top-5-ai-risks-in-the-era-of-chatgpt-and-generative-ai/?sh=717b39ec1aca
Gregory J, Kleinman Z (2023) Rishi sunak says AI has threats and risks—but outlines its potential. BBC News. https://www.bbc.com/news/uk-67225158
Grimmer J, Stewart BM (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Political Anal 21(3):267–297. https://doi.org/10.1093/pan/mps028
Günther E, Quandt T (2016) Word counts and topic models. Digital J 4(1):75–88. https://doi.org/10.1080/21670811.2015.1093270
Hagendorff T (2022) Blind spots in AI ethics. AI Ethics 2(4):851–867. https://doi.org/10.1007/s43681-021-00122-8
Hagendorff T, Bossert LN, Tse YF, Singer P (2023) Speciesist bias in AI: how AI applications perpetuate discrimination and unfair outcomes against animals. AI Ethics 3(3):717–734. https://doi.org/10.1007/s43681-022-00199-9
Hagendorff T (2024) Mapping the ethics of generative AI: a comprehensive scoping review. Minds & Machines 34:39. https://doi.org/10.1007/s11023-024-09694-w
Hale E (2023) AI could cause nuclear-level disaster, third of experts tell poll. Al Jazeera. https://www.aljazeera.com/economy/2023/4/14/ai-could-cause-nuclear-level-catastrophe-third-of-experts-say
Harris J, Anthis JR (2021) The moral consideration of artificial entities: a literature review. Sci Eng Ethics 27(4):53. https://doi.org/10.1007/s11948-021-00331-8
Hase V, Mahl D, Schäfer MS, Keller TR (2021) Climate change in news media across the globe: an automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018). Glob Environ Change 70(September):102353. https://doi.org/10.1016/j.gloenvcha.2021.102353
Hendrycks D, Mazeika M (2022) X-risk analysis for AI research. Preprint at https://doi.org/10.48550/arXiv.2206.05862
Hendrycks D, Mazeika M, Woodside T (2023) An overview of catastrophic AI risks. Center for AI Safety
Humphreys A, Wang RJ-H (2018) Automated text analysis for consumer research. J Consum Res 44(6):1274–1306. https://doi.org/10.1093/jcr/ucx104
Hunt T (2023) Here's why AI may be extremely dangerous—whether it's conscious or not. Scientific American. https://www.scientificamerican.com/article/heres-why-ai-may-be-extremely-dangerous-whether-its-conscious-or-not/
Ienca M (2023) On artificial intelligence and manipulation. Topoi 42(3):833–842. https://doi.org/10.1007/s11245-023-09940-3
Isik Ö, Joshi A, Goutas L (2024) 4 Types of gen AI risk and how to mitigate them. Harvard Business Review. https://hbr.org/2024/05/4-types-of-gen-ai-risk-and-how-to-mitigate-them
James M, Scott K (2008) Robots & rights: will artificial intelligence change the meaning of human rights. In: People Power for the Third Millennium: Technology, Democracy and Human Rights, Symposium Series, Vol 8. Bio-Centre, London, UK
Jia K, Zhang N (2022) Categorization and eccentricity of AI risks: a comparative study of the global AI guidelines. Electron Mark 32(1):59–71. https://doi.org/10.1007/s12525-021-00480-5
Jones A (2024) What risks does AI pose? BlueDot Impact. https://aisafetyfundamentals.com/blog/ai-risks/
Kaminski ME (2022) Regulating the risks of AI. SSRN Scholarly Paper, Rochester. https://doi.org/10.2139/ssrn.4195066
Kaur G (2024) Dangers of AI: exploring the risks and threats. Cointelegraph. https://cointelegraph.com/learn/dangers-of-artificial-intelligence
Kilian KA, Ventura CJ, Bailey MM (2023) Examining the differential risk from high-level artificial intelligence and the question of control. Futures 151(August):103182. https://doi.org/10.1016/j.futures.2023.103182
Kundu R (2024) AI risks: exploring the critical challenges of artificial intelligence. Lakera. https://www.lakera.ai/blog/risks-of-ai
Leach S, Sutton RM, Dhont K, Douglas KM, Bergström ZM (2023) Changing minds about minds: evidence that people are too sceptical about animal sentience. Cognition 230(January):105263. https://doi.org/10.1016/j.cognition.2022.105263
Ligozat A-L, Lefevre J, Bugeau A, Combaz J (2022) Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. Sustainability 14(9):5172. https://doi.org/10.3390/su14095172
Lin Z (2024) Beyond principlism: practical strategies for ethical AI use in research practices. AI and Ethics. https://doi.org/10.1007/s43681-024-00585-5
Lind F, Meltzer CE (2021) Now you see me, now you don't: applying automated content analysis to track migrant women's salience in German news. Feminist Media Stud 21(6):923–940. https://doi.org/10.1080/14680777.2020.1713840
Low P (2012) The Cambridge declaration on consciousness. In: Panksepp J, Reiss D, Edelman D, Van Swinderen B, Low P, Koch C (eds) Proceedings of the Francis Crick memorial conference. University of Cambridge, Cambridge
Mahrt M, Scharkow M (2013) The value of big data in digital media research. J Broadcasting Electron Media 57(1):20–33. https://doi.org/10.1080/08838151.2012.761700
Marr B (2023) The 15 biggest risks of artificial intelligence. Forbes. https://www.forbes.com/sites/bernardmarr/2023/06/02/the-15-biggest-risks-of-artificial-intelligence/
McCallum S, Vallance C, Clarke J (2023) What is AI, how does it work and what can it be used for? BBC News. https://www.bbc.com/news/technology-65855333
McLuhan M (1994) Understanding media: the extensions of man. MIT Press, London
Metz C (2023) What exactly are the dangers posed by A.I.? https://www.nytimes.com/2023/05/01/technology/ai-problems-danger-chatgpt.html
Mittelstadt B, Wachter S (2023) Expert comment: no need to wait for the future, the danger of AI is already here. https://www.ox.ac.uk/news/2023-05-15-expert-comment-no-need-wait-future-danger-ai-already-here
Neri H, Cozman F (2020) The role of experts in the public perception of risk of artificial intelligence. AI SOCIETY 35(3):663–673. https://doi.org/10.1007/s00146-019-00924-9
Nguyen D (2023) How news media frame data risks in their coverage of big data and AI. Internet Policy Rev 12(2):1–30. https://doi.org/10.14763/2023.2.1708
Nguyen D, Hekman E (2022) The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation. AI & SOCIETY. https://doi.org/10.1007/s00146-022-01511-1
Nolan B (2023) These are the 3 biggest fears about AI—and here's how worried you should be about them." Business Insider. https://www.businessinsider.com/ai-biggest-fears-risk-threat-chatgpt-openai-google-2023-6

O'Neil L (2023) These women tried to warn us about AI. Rolling Stone. https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/
Owe A, Baum SD (2021) Moral consideration of nonhumans in the ethics of artificial intelligence. AI Ethics 1(4):517–528. https://doi.org/10.1007/s43681-021-00065-0
Pałka P (2023) AI, consumers & psychological harm. SSRN Scholarly Paper. Social Science Research Network, Rochester. https://papers.ssrn.com/abstract=4564997
Palmer C, McShane K, Sandler R (2014) Environmental ethics. Annu Rev Environ Resour 39(1):419–442. https://doi.org/10.1146/annurev-environ-121112-094434
Patterson TE, Donsbagh W (1996) News decisions: journalists as partisan actors. Political Commun 13(4):455–468. https://doi.org/10.1080/10584609.1996.9963131
Pratt MK (2024) 15 AI risks businesses must confront and how to address them. TechTarget. https://www.techtarget.com/searchenterpriseai/feature/5-AI-risks-businesses-must-confront-and-how-to-address-them
Regalbuto G, Nieto P, Messier A (2023) What are the dangers of AI? Find out why people are afraid of artificial intelligence. Fox News. https://www.foxnews.com/tech/what-dangers-find-out-why-people-afraid-artificial-intelligence
Reiff N (2023) What are the dangers of AI?—Decrypt. Decrypt Media. https://decrypt.co/resources/what-are-the-dangers-of-ai
Rigley E, Chapman A, Evers C, McNeill W (2023) Anthropocentrism and environmental wellbeing in AI ethics standards: a scoping review and discussion. AI 4(4):844–874. https://doi.org/10.3390/ai4040043
Ritchie J, J Lewis (2003) Qualitative research practice: a guide for social science students and researchers. SAGE
Roe J, Perkins M (2023) What they're not telling you about ChatGPT': exploring the discourse of AI in UK news media headlines. Humanit Soc Sci Commun 10(1):1–9. https://doi.org/10.1057/s41599-023-02282-w
Rushkoff D (2022) The medium is the message. Medium (blog). https://rushkoff.medium.com/the-medium-is-the-message-fbf83929127
Russell S, Norvig P (2020) Artificial intelligence: a modern approach. Pearson
Ryan-Mosley T (2023) It's time to talk about the real AI risks. MIT Technology Review. https://www.technologyreview.com/2023/06/12/1074449/real-ai-risks/
Salvi MR, Singh DR (2023) Artificial intelligence and human society. Int J Soc Sci Human Res. https://doi.org/10.47191/ijsshr/v6-i9-13
Sarangi S, Candlin C (2003) Categorization and explanation of risk: a discourse analytical perspective. Health Risk Soc 5(2):115–124. https://doi.org/10.1080/1369857031000123902
Schäfer MS, Painter J (2021) Climate journalism in a changing media ecosystem: assessing the production of climate change-related news around the world. WIREs Clim Change 12(1):e675. https://doi.org/10.1002/wcc.675
Schopmans HR (2022) From coded bias to existential threat: expert frames and the epistemic politics of AI governance. In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society. AIES '22. Association for Computing Machinery, New York, pp 627–40. https://doi.org/10.1145/3514094.3534161
Shanmugasundaram M, Tamilarasu A (2023) The impact of digital technology, social media, and artificial intelligence on cognitive functions: a review. Front Cogn. https://www.frontiersin.org/articles/10.3389/fcogn.2023.1203077
Sharma R (2024) AI risk categorization. In: Rohan Sharma (ed) AI and the boardroom: insights into governance, strategy, and the responsible adoption of AI. Apress, Berkeley, pp 275–286. https://doi.org/10.1007/979-8-8688-0796-1_22
Shelby R, Rismani S, Henne K, Moon A, Rostamzadeh N, Nicholas P, Yilla-Akbari N, Gallegos J, Smart A, Garcia E, Virk G (2023) Sociotechnical harms of algorithmic systems: scoping a taxonomy for harm reduction. In: Proc 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). ACM, New York, NY, USA, pp 723–741. https://doi.org/10.1145/3600211.3604673
Shrum L (2017) Cultivation theory: effects and underlying processes. Wiley. https://doi.org/10.1002/9781118783764.wbieme0040
Singer P, Tse YF (2023) AI ethics: the case for including animals. AI Ethics 3(2):539–551. https://doi.org/10.1007/s43681-022-00187-z
Slattery P, Alexander KS, Grundy EAC, Graham J, Noetel M, Uuk R, Dao J, Pour S, Casper S, Thompson N (2024) The AI risk repository: a comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. Preprint at https://doi.org/10.48550/arXiv.2408.12622
Sodha S (2023) AI promises incredible benefits, but also terrible risks. It's not Luddism to rein it in. The Observer. https://www.theguardian.com/commentisfree/2023/oct/29/artificial-intelligence-safeguards-risks-ai
Sotala K, Gloor L (2017) Superintelligence as a cause or cure for risks of astronomical suffering. Informatica. https://www.informatica.si/index.php/informatica/article/view/1877
Sun S, Zhai Y, Shen B, Chen Y (2020) Newspaper coverage of artificial intelligence: a perspective of emerging technologies. Telemat Inform 53:101433
Takeshita M, Rzepka R (2024) Speciesism in natural language processing research. AI and Ethics. https://doi.org/10.1007/s43681-024-00606-3
Tegmark M (2018) Life 3.0: being human in the age of artificial intelligence. Penguin Books, New York
Thomas C, Roberts H, Mökander J, Tsamados A, Taddeo M, Floridi L (2024) The case for a broader approach to AI assurance: addressing 'Hidden' harms in the development of artificial intelligence. AI & SOCIETY. https://doi.org/10.1007/s00146-024-01950-y
Thomas M (2023) 12 Risks and dangers of artificial intelligence (AI). Built In. https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence
Turchin A, Denkenberger D (2020) Classification of global catastrophic risks connected with artificial intelligence. AI Soc 35(1):147–163. https://doi.org/10.1007/s00146-018-0845-5
Turchin A, Denkenberger D, Green B (2019) Global solutions vs. local solutions for the AI safety problem. Big Data Cogn Comput 3(February):16. https://doi.org/10.3390/bdcc3010016
United Kingdom Government (2022) Animal Welfare (Sentience) Act 2022: Chapter 22. https://www.legislation.gov.uk/ukpga/2022/22/contents/enacted
Wai J (2024) The 12 greatest dangers of AI. Forbes. https://www.forbes.com/sites/jonathanwai/2024/10/09/the-12-greatest-dangers-of-ai/
Waters R (2023) The AI revolution's slow first year. In: Waters R (ed) OPENAI's launch of CHATGPT was heralded as the dawn of a new age. But given the inherent shortcomings of generative AI, companies are wondering how useful the technology will really be. The Financial Times, pp 6–6
Weidinger L, Uesato J, Rauh M, Griffin C, Huang P-S, Mellor J, Glaese A et al. (2022) Taxonomy of risks posed by language models. In: 2022 ACM conference on fairness, accountability, and transparency. ACM, Seoul Republic of Korea, pp 214–29. https://doi.org/10.1145/3531146.3533088
Xian L, Li L, Xu Y, Zhang BZ, Hemphill L (2024) Landscape of generative AI in global news: topics, sentiments, and spatiotemporal analysis. Preprint at https://doi.org/10.48550/arXiv.2401.08899
Yampolskiy RV (2016) Taxonomy of pathways to dangerous artificial intelligence. In: AAAI Workshop on AI, Ethics, and Society, Feb 2016, pp 143–148
Zamith R, Lewis SC (2015) Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. Ann Am Acad Polit Soc Sci 659(1):307–318
Zeng Y, Klyman K, Zhou A, Yang Y, Pan M, Jia R, Song D, Liang P, Li B (2024) AI risk categorization decoded (AIR 2024): from government regulations to corporate policies. Preprint at https://doi.org/10.48550/arXiv.2406.17864
Zhou Y, Moy P (2007) Parsing framing processes: the interplay between online public opinion and media coverage. J Commun 57(1):79–98. https://doi.org/10.1111/j.0021-9916.2007.00330.x
Ziesche S (2021) AI ethics and value alignment for nonhuman animals. Philosophies 6(2):31. https://doi.org/10.3390/philosophies6020031

Acknowledgements
I am deeply grateful to Alexandra Karakas for her insightful feedback and unwavering support. I also wish to thank Mihály Héder and Imre Szabó for their constructive critique and encouragement.

Author contributions
The author is the sole contributor to this manuscript.

Funding
Open access funding provided by Budapest University of Technology and Economics.

Competing interests
The author declares no competing interests.

Informed consent
This article does not contain any studies with human participants performed by the author, and therefore, informed consent was not required.

Ethical approval
This article does not contain any studies with human participants performed by the author, and therefore, ethical approval was not required.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-04814-y.

Correspondence and requests for materials should be addressed to Marcell Sebestyén.