ORIGINAL RESEARCH

AI going rogue? An integrative narrative review of the tacit assumptions underlying existential AI-risks

Jascha Bareis[1,2] · Clemens Ackerl[2] · Reinhard Heil[2]

Received: 20 August 2025 / Accepted: 25 November 2025

Abstract
This paper presents an integrative narrative review of the tacit background assumptions underlying AI existential risk (X-risks) futures. Once confined to science fiction, concerns about AI X-risks now shape debates at the crossroads of the tech world, NGOs, politics and (social) media. Despite growing attention, the plausibility of AI surpassing human controllability remains highly contested. Examining 81 peer-reviewed papers from Scopus and Web of Science, we find a fragmented discourse characterized by bold yet often unsubstantiated claims, including accelerationist growth models and speculative calculations of catastrophic tipping points. Anthropomorphic and speculative AI conceptualizations prevail, while interdisciplinary perspectives that consider issues of infrastructure, social agency, Big Tech power position and politics remain scarce. Delineating how these speculative tendencies are detrimental to the current regulatory need to tackle AI harms, we deduce an AI X-risk heuristic and advocate for a shift in attention from the maximum possible negative consequences to the structural and socio-technical characteristics of how AI is embedded—which are the prerequisites for any AI futures to emerge.

Keywords Artificial intelligence · Existential risk · Catastrophic risk · Epistemic uncertainty · Controllability · Integrative review

1 Introduction

As artificial intelligence (AI) and its capabilities continue to advance, a once-niche debate confined to expert circles and science fiction novels is now at the forefront of public discourse: the risk of AI surpassing human controllability and causing catastrophic consequences. Existential risk (X-risk) positions discuss AI as a potentially autonomous agent. These concerns are centred around the idea of power-seeking intelligence(s) seizing control over critical infrastructure—or even humanity as a whole.

While scenarios of a sudden hostile takeover, societal collapse and apocalyptic destruction may sound like the plot of science-fiction blockbusters, they are increasingly shaping real-world policy discussions. In 2023, amid the global frenzy surrounding ChatGPT, the American Future of Life Institute (FLI) called for an immediate six-months pause to the training of AI systems, warning of the potential dangers of creating "nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us", and questioning whether we should "risk loss of control of our civilization " [36]. This appeal for a moratorium on AI development was soon accompanied by a statement from the San Francisco based Center for AI Safety (CAIS), which gathered signatures from AI experts, the broader scientific audience, and public figures. Their statement urged global leaders to recognize AI-related extinction risks—and their mitigation—as priorities on par with other societal-scale threats such as pandemics and nuclear war [24].

The narrative of catastrophic AI is already influencing power dynamics, materializing in policy decisions and creating political lock-in effects. In September 2023, European Union Commission President Ursula von der Leyen literally

✉ Jascha Bareis
jascha.bareis@unifr.ch

Clemens Ackerl
clemens.ackerl@kit.edu

Reinhard Heil
reinhard.heil@kit.edu

[1] University of Fribourg, Fribourg, Switzerland

[2] Karlsruhe Institute of Technology, Karlsruhe, Germany

Published online: 09 February 2026

echoed the warning by CAIS about the "risk of extinction from AI" in her speech to the EU parliament [100]. A year later, in October 2024, computer scientist as well as influential X-risk and AI safety spokesperson Yoshua Bengio was appointed to chair one of the European Commission's Code of Practice working groups on general purpose AI [104]. In the United States, the X-risk narrative has also begun to shape policy. The Californian bill SB1047,[1] co-drafted by CAIS, proposed the creation of an oversight board, mandatory safety testing for AI models, and legal liability for Big Tech companies. However, the bill, which was regarded as a prototype for national legislation, was framed in a distinctly alarmist X-risk narrative, holding companies responsible for "mitigat[ing] the risk of catastrophic harms from AI models so advanced that they are not yet known to exist" [13, p. 1]. Critics argue that such legislative proposals have steered and continue to steer the American policy debate toward speculative future threats rather than addressing the tangible, immediate challenges posed by AI and its societal integration [1, 67]. Perhaps the most striking example of X-risk thinking intersecting with political power is the alliance between tech billionaire and FLI external advisor Elon Musk—who already in 2014 called AI "our biggest existential threat" and that "humanity risks 'summoning a demon'" [38]—and US president Donald Trump. In a similar manner, the close ties between venture capitalist Peter Thiel, tech accelerationist Marc Andreessen and Vice President JD Vance have helped in moving the AI X-risk agenda into the center of US policymaking.

Amidst the influence of the X-risk discourse on politics and regulation, public controversies about AI scenarios beyond human controllability show a divided and polarized picture, reacting to the particularly emotional and drastic frame of X-risks. After all, doomsday scenarios imply drastic opportunity costs and a closing window of action to mitigate the absolute worst-case effects for humanity. Consequently, they are highly performative and resemble the characteristics of hypes [12], catching substantial media attention. Nordmann [73, p. 32] refers to these 'if and then' scenarios as evocative 'speculative ethics', which can create forceful and unchecked futures by means of a "radical foreshortening of the conditional". Hence, for some figures, the risks of AI are on par with other catastrophic risks such as the global nuclear threat, and thus urgently call for related policy intervention. Among these proponents are notable and well-acclaimed scientists, which grants this trajectory considerable legitimacy and media attention.[2] Conversely, critics and sceptics contend that the discussion surrounding the potential loss of AI controllability allocates important intellectual, regulatory and material resources that could be better used elsewhere, pointing to the risks and associated socio-structural issues with AI in the here and now. Examples include the monopolisation of Big Tech power, societal dependencies on privately held infrastructure, or the policing of vulnerable groups supported by AI-based means (see, e.g., [55, 69, 85, 102]). These authors discuss the mere assumption of the possibility of AI beyond human controllability as scientifically unfounded, or even as a discourse-political maneuver on the part of proponents of human enhancement, tech-accelerationism and longtermist thinking [17, 37]. Further, explicit normative critiques contest the Western mechanistic and reductionist framing of intelligence as posited by tech-accelerationist perspectives, delineating these views as not doing justice to the variety of intelligence conceptions stemming from other epistemic communities [12].

Scientific scholarship has started to investigate and structure the scattered AI X-risk discourse. There are comprehensive reviews of the academic literature describing scenarios of AI beyond human controllability and issues of AI safety, ranging from a sudden takeover by power-seeking AI, to AI runaway dynamics along a misalignment with human values [46, 47, 65, 66, 98]. Gebru and Torres [37] also offer a normative mapping of the different ideological origins of the X-risk discourse, outlining what they refer to as the TESCREAL bundle (Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalism, Effective Altruism, and Longtermism) along their respective historical roots. However, while the existing literature outlines various scenarios of AI beyond controllability, maps AI safety issues, and clusters their proponents according to their underlying ideologies, it offers little reflection on the very *preconditions* of (proclaiming) these futures.

Therefore, this paper addresses the following question: What conceptual enablers, probability horizons and tacit assumptions serve as building blocks in the construction and narratives of these futures? Essentially, despite the evocative framework and normative debates, scientific research doesn't scrutinise the underlying *plausibility* of the futures employed to the extent necessary. The perspective on plausibility matters greatly because it (de-)legitimizes political attention to certain futures and their adjunct risks. For regulators, the question of whether X-risks are what Brock [20] calls "wishful worries", i.e., "problems that it would be nice to have, as opposed to the actual agonies of the present", or plausible (even if distant) futures that they need to consider, is crucial.

As shown above, many non-scientific actors contribute towards the public debate on AI X-risks. Sotala &

[1] Turned down by Democrat governor Newsom in August, 2024.

[2] Notable signers of the before mentioned letters are well established scientists, as well as controversial entrepreneurs: Geoffrey Hinton, Yoshua Bengio, Nick Bostrom, Yuval Noah Hariri, Sam Altman, or Elon Musk.

Yampolskiy [88] also observe in their review of the literature on AI and X-risks, that this scattered yet polarized debate is dominated by non-peer-reviewed publications, such as blogs, books and commentaries.[3] A further issue are very influential Silicon Valley figures like Elon Musk, Peter Thiel, Marc Andreessen and Sam Altman, or controversial academics like Nick Bostrom and Toby Ord. These proponents stem from the Longtermist movement and make way for an uptake of sensationalist and alarmist claims about AI capabilities within scientific discourse, granting them (further) attention, political power, and allotment of venture capital [94, 95]. This strategic opportunism is staged in (social-) media and influences public opinion and sentiments around AI, which complicates a critical and structured analysis of the plausibility of AI X-risks. We take these circumstances as a further motivation to approach the construction of plausibility of AI X-risks with scientific scrutiny.

To do justice to this task, we turn to the academic debate to conduct an integrative narrative literature review [28, 92] of explicitly peer-reviewed academic publications on AI X-risk scenarios. We hypothesize that the stringent standards for indexing in Scopus and Web of Science (WoS) ensure that the included journals and articles are held to rigorous scrutiny, also concerning their quality of argumentation. This renders them well-suited for a critical, plausibility-focused review of AI X-risks futures and their underlying narratives. Internationally, these two databases are approached as reliable indicators for academic quality. Many scientific journals and academic institutions use these scientific databases to calculate various impact metrics like citation scores and impact factors, which are then used as standards to recruit scientific staff, attract funding and exhibit excellence. Hence, to be indexed in these databases is a prerequisite for many academic players to partake in the prestigious global ranking game.

The only work we identified that approaches the discourse in a comparable manner—focusing on the “assumptions of AGI”—is Blili-Hamelin et al. [17]. However, their contribution primarily explores the different conceptions of intelligence within the Artificial General Intelligence (AGI) discourse. Our search string, on the other hand, did not include such a constrained focus and targeted the entire discourse surrounding AI (without G) potentially acting beyond human controllability. While the work by Blili-Hamelin et al. [17] is rather conceptual—zooming-in only on a small-set of literature—it provides helpful orientation points.

We examined 81 contributions listed in Scopus and Web of Science Core (WoS) based on the following research questions:

1. How is AI defined and related to existential risk, and how is risk understood?
2. How are time, probability and plausibility horizons conceptualised concerning the risks of an out-of-control AI?
3. Which background conditions (material, institutional, economic) as well as societal circumstances are discussed in the futures towards out-of-control AI?

The paper is structured as follows. First, the conceptual and methodological approach is presented, tackling the challenge of how to assess a speculative realm of a far-fetched future and its associated scientific discourse. Then, the integrative analysis addresses the three questions above, deconstructing the prevalent conceptual background conditions and tacitly deployed narratives. The discussion leads us to propse an AI X-risk heuristic (Table 1), and the conclusion advocates for a discursive shift in attention from the maximum possible negative consequences of AI to the characteristics of AI that are assumed to be conditions for these consequences in the here and now.

2 Methods of assessing the speculative: a hermeneutical approach to unravelling tacit assumptions

This review assesses and systematises a body of knowledge that requires distinct conceptual and methodological lenses: How to assess something that is so speculative, yet would have the most detrimental consequences if it were to occur? While the former diminishes the urgency from an epistemological and political point of view, the latter highlights the high stakes involved. Existential risks differ fundamentally from other types of risks. In general, risks are always associated with uncertainty and are usually quantified according to the formula “amount of loss times probability of occurrence” to render them tangible and allow for a certain degree of comparability [78]. However, the speculative character of ‘existential’ risks, coupled with their implied ‘all-humans-affecting’ scope, makes their assessment a particularly challenging scientific undertaking.

With regard to uncertain future technology and innovation paths, Technology Assessment (TA), Responsible Research and Innovation (RRI) and Science and Technology Studies (STS), have reacted with future directed heuristics. Speculative trajectories can be deconstructed and substantiated, among others, through hermeneutical vision assessment [35, 43], approaches oriented at anticipations and expectations [4, 56], or forecasting methods [61, 62]. Scrutinizing speculative futures and their plausibility

[3] Even though there is a scientific community assembling around the conferences and Journal of the Artificial General Intelligence Society.

involves analysing hypotheses that lack the crucial epistemological quality of verifiability or falsifiability. The discourse revolving around AI X-risks is characterised by such a high degree of uncertainty that their “conus of possible futures” cannot be reduced argumentatively to a manageable number of assessable scenarios [45, p. 68], thus allowing for neither a prognostic nor a scenario-based literacy. Instead, the discourse at hand reflects a plurality of tropes and meanings, with overlaps of scientific and non-scientific views, and standpoints that mediate between fact and fiction. All in all, the AI X-risk discourse encompasses speculative futures that can neither be proven or disproven, but only be assessed along the spectrum of plausibility. This is why, in our analysis, we place particular emphasis on the *background(ed) assumptions* leveraged to lend plausibility to the respective AI future in question. This analysis establishes not only the foundation for a scientific critique of the prevalent positions, but also an identification of the “deficiencies, omissions, inaccuracies, and other problematic aspects” [92, p. 362] found in the prognostics of the AI X-risk discourse. Hence, we will not only focus on what is said, but also on what is ignored and muted because it could complicate the projection of AI futures to come. As will be shown in the analysis below, this notion of overshadowing—understood as strategic ignorance [64]—creeps into the AI X-risk futures through the negation of social and scientific complexity, or the sidelining of counter-arguments from different scientific disciplines.

Looking at the consequences sketched, the AI futures under consideration work with a certain rhetorical escalation, employing maximum stakes along the lines of ‘existential’, ‘going rogue’, ‘beyond controllability’, ‘gaining the upper hand’ and ‘apocalyptic’ framings. As many positions in the corpus emphasise the utmost urgency in dealing with existential risks, we also examine the theatrical techniques [39] and the “dramaturgical regime” of the articles in question—implying that that their “performative imaginations are enacted” through figurative language and performative tweaks that can be deconstructed [74, p. 259]. This motivates us to pay particular attention to the rhetorical devices utilised, such as metaphors, analogies and exaggerations, which underscore narratives of grandeur, predictive precision, or threat (as practised in hermeneutic approaches; e.g. in vision assessment [35], RRI [44] or in hype studies [12]).

This integrative narrative literature review is limited to original articles listed in Scopus and Wo S. The search in both databases was conducted on 27 September 2024, using the following Boolean search terms developed in accordance with our research interest: ‘(ALL=((’existential risk* ‘ OR “catastrophic* risk*” OR “human extinction”) AND (’artificial intelligence ‘ OR ai))) OR ALL=(’uncontroll* artificial intelligence ‘ OR “uncontroll* ai”) and “(ALL=(rogue artificial intelligence OR”rogue ai’))’. The following inclusion criteria were applied in the selection of articles for review: The publication primarily focuses on existential risk, particularly the role of artificial intelligence, and is written in English. These criteria were then used to evaluate the titles and abstracts of all retrieved articles. Articles that did not meet the eligibility criteria were omitted from the review, e.g. articles that only briefly mention X-risks without discussing them further. A total of 233 contributions were found (WoS 74, Scopus 159). 45 duplicates were removed. The abstracts (if available) of the remaining 188 articles were reviewed for relevance and then discussed by the author team. 110 articles remained for qualitative assessment along the questions delineated above. Each article in the final sample was read and reviewed by (at least) two of the three reviewers, based on a coding sheet informed by the research interest sketched out above. During this qualitative analysis, a further 29 articles were eliminated from the corpus given that they fall short in meeting the specified criteria.

3 Integrative narrative analysis

1. How to define AI? From performative anthropomorphisation and complexity games to sentient AI

As a departure point of our scrutiny of the X-risk discourse, we examine how the given literature defines AI and relates it to existential risk (RQ 1). The comparison with human capabilities constitutes a focal anchor for many authors seeking to define AI, next to metaphysical and contested AI conceptualization as discussed below.

3.1 Turing’s AI: setting the ground for the AI vs. human competition

The Turing test remains a popular point of reference (e.g., [1, 2, 16, 52]). The test, originally termed ‘imitation game’ by Alan Turing, is widely regarded as the founding momentum of modern AI research, placing human–machine comparison at the very centre of the debate. Since Turing, this comparison is invoked to track AI capabilities, with humans representing a competing rival. This competitive framing suggests a recurrent anthropomorphisation as a backdrop and an AI potentially out of control as one that perspectively outruns the human. To evaluate the performance, it is common practice in computer science to propose challenges or quests in which the players compete in a human vs. machine format. Historically, these challenges have been modelled in the form of complexity games, involving *breakthroughs* in the fields of chess (‘Deep Blue’), ‘Jeopardy!’ and ‘GO’ (‘Aleph Alpha’). In the literature corpus, there is a prevailing tendency to invoke benchmarks to measure the performance in complexity games, with the objective of providing a tangible illustration of AI’s capabilities [29]. Doing so, authors subtly imply that the resolution of humanity’s challenges is tantamount to a quest for mastering complexity. The apparently hardest complexity challenge that contemporary AI, in the form of Large Language Models (LLM), is facing is aptly named ‘Humanity’s Last Exam’ [75]—before even titled ‘Humanity’s Last Stand’ [80]. This framing rhetorically implies a performative, looming threat for humanity, and drastic consequences if AI ‘passes the exam’, ultimately testifying that AI can defeat and conquer the human.

Benchmarks used as a point of reference for AI capabilities have been criticized for being narrow and reductionist, essentially committing an inductive fallacy. Raji et al. [78] and Eriksson et al. [33] argue that benchmarks problematically jump from *domain* performances to conclusions about *general* AI capabilities. As Blili-Hamelin et al. [17, p. 146] out: “AI evaluation practices like benchmarking, as currently practiced, mostly treat individual models as bearers of the properties they measure, and definers of AGI often propose tests that are mostly or entirely tests on individual agents”.

Another pivotal component in delineating AI pertains to functional characteristics. This approach primarily targets the capabilities that an AI must achieve in order to be classified as intelligent (e.g. [5, 30–32]). It is important to note that this cluster does not constitute an independent category but rather represents an instantiation of what *exactly* the machine would need to be able to perform, in order to outperform humans. As Dung [31, p. 138] suggests, “[t]hese systems presumably need to be superior to humans in some strategically important domains (e.g., general planning, reasoning speed, persuasion, hacking, science etc.) and to have some notable degree of competence in many domains”. The specified functional domains that an AI must master (in order to be designated as such) are largely borrowed from engineering and information processing, as well as their respective categories and models. These include tasks such as command and control from cybernetics, brain modelling, cognitive science and other technical fields that engage with perception, memory, reasoning, learning, or language. Correspondingly, the scientific indicators, models and theorems in the examined literature draw upon the domain of technical apparatuses and their symbolic systems—such as formulae, modularity, lists, tables and actions—which are enacted as computation (see [60]). Moreover, as humans are taken up as the point of reference in the journey towards machine breakthroughs, this discourse does not hesitate to apply these technical analogies to humans—and in doing so, converting them into *machine-like entities*. This dynamic is also observable in robot-ethics, where machines become equated to humans as “quasi-others” [25, p. 75], and are discussed as moral recipients with rights [101]. Thus, we can note that this part of the examined literature frequently reflects a functionalist and engineering-based understanding of capabilities and applies it as a simile to both machines and humans, staging a competition between them.

3.2 Speculative AI out of control

A considerable proportion of the contributions examined invoke a definition of AI in the context of a loss of control, referring to the attainment of mental/cognitive states such as consciousness, awareness, sentience, autonomy, or moral sentiment (e.g., [5, 16, 30–32, 49, 53, 82, 93]). While the above depicted human–machine rivalry focused on *outcomes* in complexity games (which are operationalizable), these criteria often involve internal *states, processes* and *judgements* of the mind (thus phenomena which slip external manifestation).

For example, Beltramini [16, p. 258] “define[s] 'intelligent machine' in terms of sentient machines with general artificial intelligence ('AI')”. Similarly, Harvey [49, p. 49] stipulates that AI will develop motivations as it becomes aware of its “own mortality”. Saavedra-Rivano [82] follows this track, specifically pointing to evolutionary epochs:

> “We look at the impact of AI in the short term and the longer term, the dividing point between these two periods being the moment when AI entities would acquire self-consciousness and be able to reason according to their own views of the world. There is of course plenty of discussion on whether that state of 'sentience' is possible at all for machines and also on when that would happen” [82, p. 319].

Even if machine *sentience* is not presented here as scientifically proven, it is nevertheless invoked as a distinguishing feature of epochs, of a ‘before’ and ‘after’. However, these assumptions also serve as an indication of how highly speculative the discourse on AI beyond human controllability is in parts of the scientific literature. Another criterion brought to the fore is the achievement of autonomy, understood not only as machine self-*control* or self-*sufficiency*, but also as self-*awareness*. Arvan [60] reflects that human-like artificial moral agents would exhibit consciousness, intentionality as well as having a free will. In scenarios and conceptualizations that attribute the possibility of sentience and (self-)awareness to AIs, also an AI ‘patient ethic’ emerges as a proximate prospect. For example, Dung [30, p. 7] poses the question of “How to deal with risks of AI suffering?”—specifically addressing the potential suffering of AI, not the potential suffering of humans caused by AI (see also Beltramini [16]).

We observe that, in the literature under consideration, the reference point quickly changes from AI to AGI and singularity when mental states related to consciousness, awareness and sentience are evoked. This marks the point at which authors enter a highly speculative realm. Instead of providing evidence that (and showing *how*) AI *can* actually attain mental states like consciousness, authors tend to postulate these circumstances simply as a given. Also, problematically, AGI remains notoriously ill-defined in the corpus. Observing how authors approach AGI in their papers, we suggest that it should rather be understood as a mindset than a founded theory. It is associated with the strong and assertive belief that AI will attain some mental states that make it equal to or surpass the human. Some papers show an almost messianic confidence towards this trajectory. Here the attainment of machine consciousness and sentience is frequently posited to be a matter of time or high likelihood, rather than of categorical possibility. This implied functionalist theory of mind [21] suggests that emergent internal phenomena such as consciousness will *appear* solely as a result of advancing algorithmic performance. However, we note that other schools of thought in philosophy, psychology, neuroscience and biology have, for decades, challenged this assumption as implausible and speculative, depicting an ongoing, unresolved debate in the philosophy of mind [34, 50]. It remains unclear how phenomena of a different quality (e.g., consciousness) shall emerge by (simply) more of the same (e.g., algorithmic performance).

3.3 Contested assumptions and the lack of benefit of an AI definition

A small number of the texts in the given corpus contend that the discourse surrounding the potential loss of AI control is highly speculative and controversial—and even without value. Johnson and Verdicchio [54, p. 586] acknowledge the contradictory nature and high degree of speculation within certain positions across the discourse, stating: “The absence of any real understanding […] gives futuristic thinkers a free hand to present misleading and sometimes contradictory scenarios” (see also [59] or [91]). A final, albeit small, part of the assessed literature distances itself from the benefits of a consolidated definition of both AI and X-risks. Instead, these authors argue that the discourse should centre on the protected goods—such as human rights, safety, or human autonomy—that are potentially compromised by AI systems. The emphasis here is on the concrete potential for harm to individuals, irrespective of the complexity or the ‘intelligence’ of the algorithmic system in question. For instance, Bajgar & Horenovsky [7, p. 1048] present a focus on negative human rights, which would be “relatively agnostic to where we draw the line of what already counts as an AI system, and it could include present-day systems such as autonomous vehicles, robotic cleaners, or recommender systems”. Negative human rights, as mentioned by Bajgar & Horenovsky [7] encompass the right to life, security of the person, the right to property, and the right not to be tortured. These rights are primarily discussed within the US legal sphere as rights of non-interference.

2. When will it happen? From the calculation of time and probability horizons to a sudden intelligence explosion

The considerable uncertainty surrounding the timing, and indeed the very possibility of the emergence of AI beyond human controllability, invites a broad spectrum of forecasts and speculations. How are time, probability and plausibility horizons conceptualised concerning the risks of an out-of-control AI (RQ II)? One tendency in the literature reviewed is to predict, or even calculate, probabilities and timeframes for the occurrence of events associated with AI X-risks. This practice is often, as discussed above, linked to the emergence of AGI. The strand of literature treats the occurrence of future events as a calculable phenomenon that can be modelled and scientifically determined.

3.4 Speculative futures meet calculations of occurrences

Predictions about the time horizons for the development of AGI vary considerably among experts, to the extent that both their accuracy and their usefulness for guidance are called into question. Graham [42] points out that, although experts and researchers do not necessarily agree on an exact timeframe for the emergence of AGI, many contribute their own estimates to the discussion. Several papers (a.o. [27, 72, 76, 97]) refer to the expert survey on human-level AI by Grace et al. [41], which found that experts, on average, predict a 50% probability of the attainment of human-level AI by 2061. Dung [32] “argue[s] for the conclusion that AI will lead to the permanent disempowerment (e.g. through extinction) of humanity by, at the latest, 2100.” Other authors within the literature corpus at least consider it a significant chance that AGI will be reached by the end of this century [40, 63, 84, 107].

Many papers link the attainment of X-risk AI with the coupling of risk assessment metrics and measurements, referring to notions of safety and controllability. Goldstein and Kirk-Giannini [40, p. 1] propose a comprehensive risk assessment and speculate that the likelihood of a potential existential catastrophe resulting from misaligned AGI can be reduced to approximately 0.05% through the advancement of language agents (LLMs, e.g. ChatGPT). In their view, these agents can assist in resolving the "three important issues related to aligning AIs: reward misspecification, goal misgeneralization, and uninterpretability". Yampolskiy [107] emphasises that the majority of advanced AI initiatives are devoid of integrated safety mechanisms, thereby underscoring the probability of the first AGI system being safe as being extremely low. This view is supported by Turchin [96, p. 47], who proposes the establishment of a minimum acceptable level of AI risk as 5% of the "cumulative probability of powerful AI's creation in a given time frame". Moreover, Turchin [96, p. 47] estimates that there is a 5% probability of AGI development within the next few years, coupled with the assumption that "human-level AI will pose an existential risk as soon as it will be created".

We note that such probability, threshold and occurrence calculations are very widespread in the literature and, as emphasised above, point to a peculiar reductionist worldview within the X-risk community. From this perspective, events and occurrences, no matter their complexity, can be formalized, calculated and modelled. We propose that these predictive practices can be understood as 'technologies of distance' [77] that have an important psychological effect. From this perspective, numerical models and probabilistic estimates carry an aura of perceived objectivity and trustworthiness. In contrast to careful descriptive accounts that acknowledge their limitations or to future-directed scenario building, which remains subject to pathway interpretation and plausibility assessments, numerical calculations in relation to future(s) convey a sense of precision, truth, and impartiality. As such, they confer authority to those who employ them, rendering them into "experts" capable of predicting—or even calculating—the future. Moreover, we suggest that the use of metrics tied to future events is performative for the entire AI future discourse: If a claim suggests that the definite occurrence x can be reduced by y% through measurement z, the consequence is an implicit imperative to take action.

The tendency to quantify risks and score probability of occurrences enables the mobilisation of a questionable ethical framework—'longtermism'—as a theoretical backdrop in parts of the discourse on AI X-risks. Not to be confused with long-term thinking, longtermism posits that the lives of all possible future human beings are more valuable than those currently living on Earth. This, in turn, implies a moral obligation to maximise the total amount of 'value' in the universe and ensure our species does not fall short of its potential (see [95]). A key figure of reference in our corpus is the philosopher Nick Bostrom [19]. His conceptualization of AI X-risks is grounded in their potential impact on the advancement of 'humanity's potential', including the moral value of all potential future generations of mankind. This highly speculative reasoning, which ties moral imperatives to a far future, is taken up by some authors in the literature reviewed. For example, in "Human Extinction and AI: What We Can Learn from the Ultimate Threat", Lavazza and Vilaca [58] advocate for the creation of 'heirs'—"humanoid robots that reproduce our salient characteristics by imitation, thanks to AI powered by machine learning". They assert that the potential extinction of Homo sapiens necessitates proactive measures: "It might be worth starting to contemplate a way to pass what is the best in the human species to what might be our 'heirs'." Attributing moral empathy to a speculative future is not met without opposition in the corpus examined. In "The Problem with Longtermism", Hyde [51, p. 149] rejects longtermism outright, calling it "absurd" and dismissing its central premise—"the idea that we can even be morally concerned about what is a million years away, yet again obliged to do something about it"—as "utter folly".

Also Thorstad [91] presents a rebuttal of the underlying assumptions supporting the calculation of the "AGI breakthrough". He deconstructs models and theorems of accelerating growth, pointing to bottlenecks and physical constraints that, from his point of view, most X-risk scholars simply fail to consider. For Thorstad, the singularity hypothesis rests on growth assumptions that are not supported by current empirical observations, noting the coming to an end of Moore's law with a declining pace of hardware growth and diminishing research productivity. He concludes with a call for scientific humility and robustness:

> "The singularity hypothesis posits a sustained period of exponential or hyperbolic growth in the intelligence of artificial agents, continuing at least until machines exceed humans in intelligence by as much as humans exceed mice. These are extraordinary claims, and they should require correspondingly extraordinary evidence" [91, p. 4].

3.5 Sudden loss of control

Some scholars highlight the possibility of a sudden emergence or uncontrollable acceleration of developments leading to AI surpassing human controllability. Torres [93, p. 105], for instance, discusses the risk of AGI emerging out of the blue and in an unpredictable manner: "AGI could be a momentary flash between sub-human-level AI and artificial superintelligence. An artificial superintelligence is a system that could significantly outperform every possible human in all cognitive domains". Jilk [53, p. 429] depicts a similar pathway: "In an intelligence explosion, initial creation of artificial intelligence with a critical mass of capabilities and drives is followed by an inexorable process of increases in that intelligence. […] This process is usually viewed as uncontrolled, unstoppable, and accelerating". Others question this run-away-dynamic. In their review of the X-risk literature, Armstrong and Sotala [5] stress the inherent uncertainty in predicting the development of advanced AI. Likewise, Yudkowsky [109] argues that pinpointing the emergence of AGI is nearly impossible. He suggests that clear evidence of AGI on the way may only become apparent within two years of its arrival ([109], see also [85]).

We argue that the narrative of a sudden emergence of AGI carries psychological implications. Irrespective of intentionality, it paints a picture of looming threat, suggesting that humanity has merely a limited *kairos* window of action before reaching the point of no return. In the examined literature, the notion of a disruptive, accelerationist technological development is frequently employed as a rhetorical device to symbolise an unpredictable and unprecedented moment in which a superintelligence bursts upon humanity and *invades* or *overthrows* it. However, the discourse is not clear concerning the indicators to look out for on the horizon regarding potentially threatening developments. Hanson & Yudkowsky [48], for example, argue that the development of advanced AI does not depend primarily on hardware availability, but on the emergence of an unprecedented seminal concept (i.e., an unforeseeable genius momentum). The probability of such an occurrence, as posited by Turchin [96], is seen as heavily influenced by the number of researchers engaged in AI development. In this view, the likelihood of such a pivotal conceptual leap is closely associated with the financial and attentional resources allocated to AI research at a given time. Although psychologically powerful, we believe that this narrative necessitates a careful plausibility check regarding its economic, infrastructural and material prerequisites—which leads us to the last research question.

3. Unencumbered by socio-material constraints: On what is ignored

Any firm declaration concerning (far-fetched) AI futures is contingent on a wide array of prerequisites. AI does not develop in isolation but is embedded in the social and material, ranging from (geo-)political climate, economic investments, to the access to chips, training data and server farms. Given the aim of assessing the plausibility of proclaimed X-AI futures, these domains are pivotal because they carve out the very resource and infrastructure any AI development depends on. This leads us to ask: Which background conditions (material, institutional, economic) as well as societal circumstances are discussed in the futures towards out-of-control AI (RQ III)?

Here the picture among the assessed literature is striking. We observe that almost no contribution in the corpus does justice to mention any background conditions. Instead, AI development is approached as a stand-alone 'autonomous' agent, whose steady increase in capabilities is just depicted as a given, pointing to a technological determinist worldview [104]. In consequence, AI is sheltered inside a closed, formalized world which is imminently theory and model based. Examining our corpus, we note that the AI X-risk discourse seems predominantly popular in the natural sciences, lacking interdisciplinary papers that incorporate insights from the humanities and social sciences. This may be the very reason for the dominance of functionalist and engineering parameters over social ones in the corpus. Recalling the notions of strategic ignorance and overshadowing from the introduction, it is revealing which aspects the X-risk discourse does *not* discuss in order to enable the postulation of catastrophic trajectories. The disregard for counterarguments, lessons learned from history, and insights from other scientific disciplines is very salient within the examined literature.

A few reflective exceptions exist. Within our literature corpus, Singler [87, p. 174] provides a critical analysis of the prevailing narratives surrounding "AI apocalypticism", contending that the (anticipated) advancements of AI "will, and do, also intermingle with our dreams of the future and affect our conceptions of ourselves now […] depending on the narratives around them that are generated by both lay and expert audiences". Singler's note is similar to approaches in the social sciences which argue that AI should be considered situated and relational [60, 89, 90], reworked and understood by different users and narratives, and enmeshed in constellations of power.

Moreover, Thorstad [91, p. 10], already mentioned above, points to concrete bottlenecks and limitations regarding the resources and infrastructure AI depends on. He questions the accelerating growth of AI capabilities along the limits of producing ever smaller and denser transistors, which makes both their manufacture and interaction increasingly complex: "Any viable path to improving artificial intelligence will eventually bump up against resource constraints and the laws of physics in ways that are not easily overcome". Moreover, our corpus includes a paper that explicitly addresses the strategic interests and power position of major technology companies within the domain and discourse of X-risks and related research. Leggett [59] adopts a critical power perspective and reflects on the societal preconditions of an out-of-control AI. By establishing an analogy between AI beyond human controllability and the capitalist economic system with its accelerating growth dynamics, he challenges conventional assumptions and offers a novel perspective on the matter:

> "Superintelligence? In short, we have created a corporate market machine that is now capable of manipulating and controlling individual humans, and that is infinitely better, already, at this task than any human is, or could hope to be. And we have given this machine

> the single, overarching goal of obtaining a return to capital" [59, p. 736].

The commentary cautions against a debate that disregards systemic risks for both individuals and society as a whole. It points to the power wielded by a privileged group of affluent tech oligarchs, the ownership over AI infrastructure, as well as to the capitalist and sensational driven logics of media attention.

Overall, such critical voices represent the stark exception rather than the norm within the examined literature. Consequently, the reviewed literature exhibits insufficient awareness of the constraints imposed by the socio-material realm, frequently containing bold predictions, questionable calculations, a functionalist theory of the mind, and an unwavering—at times, quasi-messianic—conviction in the AI singularity trajectory.

4 Discussion of the results

The perception that AI development is accelerating beyond human controllability has fuelled a proliferation of narratives utilising bold hypotheses and strong doomsday rhetoric—even in peer-reviewed journals indexed in core databases such as Scopus and WoS. The results of our integrative narrative analysis are summarised in the table below (Table 1). This table further features an AI X-risk heuristic, aiming at deconstructing and identifying problematic AI X-risk conceptualizations and doomsday narratives. With this, we hope to steer the discussion away from anticipating the most extreme and catastrophic outcomes of AI—which devotes primary attention to preventing these hypothetical scenarios—and advocate for both a de-escalation of rhetoric and a shift in focus to the here and now. We propose a toolkit for analysts, policy makers and academics to substantiate the speculative, emotionally charged, and hype-fuelled debate about AI futures.

The question remains: what role shall speculation play in assessing vast—but possibly very detrimental—futures linked to technological developments? Speculatory philosophical arguments or anticipatory future games play an important role because they stimulate imagination and out of the box thinking. Breaking out of established thought patterns and cognitive path dependencies can open the gates to new future trajectories—even if at times vast and disordered. Hence, is the problem the practice of speculative accounts based on shaky premises, or the very immediate impact these speculations have on contemporary policy?

Here, the issue any speculatory account has to consider is that philosophical speculation never happens in a social vacuum. The AGI debate is not a purely academic discourse but has a massive impact on society. As outlined in the introduction, powerful corporate players hype up the AGI narrative to bolster the value of their companies. These companies are among the most powerful contemporary actors around the globe. They hold the reins to steer the hype machine, owning social media platforms, research units, and providing LLM chatbots which increasingly take over societal "knowledge" production. If they use the frame of AI-out-of-control while being the very developers and providers of AI, attention from technologists, politicians, journalists, and the general public is certain.

Further, what makes the AGI debate rhetorically so influential is its story. Essentially, it offers a highly emotional and dramatic plot that could end in a tragedy: Hyperboles such as 'existential', 'going rogue', 'gaining the upper hand' and 'apocalyptic' are evocative and carve out an attention grabbing framing. It is not surprising that the media and politics react to these outcry scenarios. Thus, speculating responsibly means to be aware of this social setting and rhetoric framing. Attention resources of politics and general public, especially in an age of social media with constant stimulation and news overload, are limited. Any resources spent on a particular future trajectory inevitably implies a redirecting of resources from another potential future or present issue of general interest. Rhetoric like dramatization can be a blessing (for grabbing attention)—but also a curse (if distracting from other important issues).

The aspect of relying on strong or weak premises in constructing futures is another important pillar in this debate. Working with robust premises as a baseline to then discuss speculative futures, inevitably ties fantasy or logical coherence back to empirical reality: the stronger the premises, the more convincing and plausible the constructed futures. It is quite telling that none of the most limiting contemporary boundary baselines like geopolitical chip wars, financial bubbles, inflation, climate change costs, right wing populism, the rise of the African continent or Chinese sphere of influence with their respective vision of AI, or the centralization of AI infrastructure, made it into starting conditions to speculate about AI (X-risk) futures. Speculating about AI as sheltered inside a closed, formalized world which is imminently theory and model based can be a fun and wild task, but if such speculation really shall support *future literacy,*—and in so doing, provide orientation for action—it must itself be linked back to empirics and given socio-material constraints. This is also the best protection against instrumentalization from attention-seeking actors, who may appropriate the cause for economic or socio-political goals. One can easily manipulate and steer futures, as by their very nature, they transcend into the not-yet-experienced—but to negate the empirical reality they depend on is a far more difficult task.

Table 1 Results and AI X-risk heuristic

Research question	Results of the review	AI X-risk heuristic
Variable of comparison: How is AI defined? How is it linked to existential risk? (RQ I)	Many papers in the X-risk community draw a human–machine analogy, displaying a tendency towards anthropomorphisation and a mechanistic worldview. AI in competition with humans—the goal is to surpass humans.	Are the progress and breakthroughs of AI justified by closed-world games, challenges, quests and benchmarks?
	Benchmarking as an indicator of AI performance with the problematic tendency of inductive generalization.	How prevalent are anthropomorphic tendencies, such as equating human capabilities and understandings with those of machines?
	AI defined via functional capabilities, implying an engineering understanding. Few papers: The value of defining AI is contested. Lack of benefit, rather focus on protected goods as negative human rights.	Which metaphors or symbolic references are used to describe the relationship between humans and machines?
Speculative assumptions in relation to AI and (loss of) control (RQ I)	A considerable proportion of the literature invokes speculative criteria such as consciousness, sentience or autonomy.	Are speculative criteria mobilised to define AI? Does the referent change from AI to AGI?
	High confidence that AI will attain these criteria, with AI developing into AGI/Superintelligence.	Is AGI delineated as a point of reference? How, with what reference points? If so, by what criteria does AI develop into AGI?
	AGI and its development paths are ill-defined. Advocation for functionalism as a theory of mind, treating AGI as an emergent process in the machine.	What is suggested as a proof that AGI is reached and that humanity loses controllability?
	Some papers suggest a moral patient ethic, ascribing ethical value to machines.	Is AI's attainment of (super-)human properties seen as inevitable / an emergent result of a mere matter of more algorithmic power?
Indicators of development (RQ II)	Calculations of probability, thresholds and occurrences are very widespread in the AI X-risk community. Numerical models produce an aura of perceived objectivity, precision and truth. This bestows trust and changes the role from 'authors' to 'experts'. The use of metrics evokes a sense of certainty and triggers a performative notion, calling for action.	Is the growth in AI capacities projected to be linear, exponential or contingent?
	Some scholars proclaim the likelihood of a sudden loss of control ('AGI flash'). Viewed as uncontrolled, unstoppable, and accelerating.	Are X-risks presented as calculable events that can be modelled and scientifically determined through probability-, threshold- and occurrence- calculations?
	The effect is performative, suggesting a small window (*kairos*) to react.	What is the tone of presentation, confidence, imagery used? Are rhetorical devices utilised to convey rupture, breakthrough, acceleration?
Socio-political and -material embedding of X-risk future (RQ III)	Almost no contribution present. Instead, AI is portrayed as an 'autonomous agent'.	Are the arguments surrounding X-risks embedded in economic, material and socio-political factors? Any questions of ownership, infrastructure or regulation raised?
	Predominant X-risk discourse lacks interdisciplinary approaches and insights. This results in an overshadowing of the social and material, liberating AI development from any real-world constraints.	Is the development of AI argued as imminently theory/model based?
Who's speaking? And for whom? (RQ III)	Mostly computer scientists and engineers. Philosophers like Nick Bostrom as prominent figures of reference, 'Longtermism' as a prominent school of thought.	Do the authors, and/or their organizations, have a distinct / (ill-)legitimate interest (personal, financial, political), in drawing attention to AI development?

5 Conclusion

In this paper, we interrogated the peer-reviewed corpus indexed in Scopus and Web of Science (WoS) for the plausibility conditions of AI X-risks. We hypothesized that the stringent standards for academic indexing ensure that the included journals and articles are held to rigorous scrutiny, also concerning their quality of argumentation. The result reveals a multifaceted, and somewhat disconcerting picture. Returning to the introduction, the proclaimed AI X-risk futures can rather be labelled as speculative “wishful-worries” [20], than plausible trajectories that policymakers must take into account in a prioritized manner.

In approaching and defining AI (RQ I), a significant portion of authors privilege alarmist narratives that rest on anthropomorphic conceptualisations of AI and a functionalist theory of mind—some even attribute faculties such as ‘consciousness’, ‘autonomy’ and ‘sentience’ to computational systems. This framing not only imbues the scientific discourse with emotional, speculative expectations and in so doing, undermines its analytical value. Moreover, the navigation of these categories quickly comes along with a new background condition: a jump from assessing Artificial Intelligence to a much more speculative object of investigation, namely the attainment of Artificial General Intelligence—often used exchangeably with terms as singularity and superintelligence—and the accompanying risks of such development. Both this jump and the related concepts, however, remain vaguely outlined and ill-defined within the corpus at hand.

Concerning time, probability and plausibility horizons towards an AI out of control (RQ II), we observe a salient tendency in the corpus. AI future events and occurrences, no matter their complexity, are approached as phenomena that can be formalized, calculated and modelled. We note that when authors use these numerical models, they produce an aura of perceived objectivity, precision and truth. This bestows trust on academic authors and changes their role from ‘authors’ to ‘experts’. Some scholars also proclaim the likelihood of a sudden loss of control (‘AGI flash’). This development is presented as uncontrolled, unstoppable, and accelerating. However, there are also a few authors rejecting these premises as implausible, pointing to resource bottlenecks, a declining pace of hardware growth and diminishing research productivity.

Regarding questions of background conditions (RQ III), authors often overlook the socio-technical foundations of AI, dedicating minimal attention to the necessary infrastructural, political and material preconditions of a supposedly accelerating AI towards loss of control. The predominant X-risk discourse lacks interdisciplinary approaches and is rather dominated by computer scientists and some analytical philosophers. Critical voices, which do exist, seem to publish elsewhere, but not in the world of WoS and Scopus.

This also shows a limitation of this study. Retrieving scientific literature from impactful scientific databases covers substantial peer-reviewed literature—but it does not cover all publications, and seemingly, not all *scientific views* in the field.

These insights hint to a greater dimension that can only be touched upon briefly here: are impact factors and citation scores undermining diversity in scientific debates, sidelining more carefully, critically crafted scholarship taking place at smaller and/or less established scientific journals? To what extent is the publishing business, which is ever more building on metrification systems, vulnerable to sensationalist scientific claims designed to boost citations? These aspects may play a particular role in the contexts of the significant promises and fears projected at (the development) of AI.

The question to what extent the AI hype has an impact on publishing culture and also overall scientific authority is a salient one—given that many notable computer scientists who have delivered great achievements in their discipline (take Geoffrey Hinton who just recently won the Nobel price in Physics or Yoshua Bengio) are the ones who make strongest claims on the looming societal threats or on the paradisiac potentials around AI.

Acknowledgements The authors would like to thank Isa Luiten who prepared the manuscript for publication.

Author contributions All authors whose names appear on the submission made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data used in the work; drafted the work or revised it critically for important intellectual content; approved the version to be published; and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The order of authorship was agreed upon and approved by the author team internally. We declare the contribution of the first author (Jascha Bareis) and the second author (Clemens Ackerl) as being equal.

Funding Open Access funding enabled and organized by Projekt DEAL. The work was funded by the German Federal Ministry of Research, Technology and Space (16IS23075).

Data availability The full list of the literature corpus can be accessed under: https://zenodo.org/records/17865743.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Abney, K. A.: Space War and AI. In: Masakowski, Y. (ed.), Artificial Intelligence and Global Security: Future Trends, Threats and Considerations, pp. 63–79. Emerald Publishing (2020)
2. Alfonseca, M., Cebrian, M., Anta, A.F., Coviello, L., Abeliuk, A., Rahwan, I.: Superintelligence cannot be contained: lessons from computability theory. J. Artif. Intell. Res. **70**, 65–76 (2021). https://doi.org/10.1613/jair.1.12202
3. Altman, S.: Planning for AGI and beyond. OpenAI. https://openai.com/index/planning-for-agi-and-beyond/ (2023). Accessed 13 August, 2025
4. Alvial-Palavicino, C.: The Future as Practice. A Framework to Understand Anticipation in Science and Technology. Tecnoscienza – Italian Journal of Science & Technology Studies 6(2) (2015). https://doi.org/10.6092/issn.2038-3460/17262
5. Armstrong, S., & Sotala, K.: How we're predicting AI–or failing to. In J. Romportl, P. Irving, E. Zackova, M. Polak, & R. Schuster (eds.): Beyond artificial intelligence: The disappearing human-machine divide, pp. 11–29. University of West Bohemia (2015)
6. Arvan, M.: Varieties of artificial moral agency and the new control problem. Humana Mente **15**(42), 225–256 (2022)
7. Bajgar, O., Horenovsky, J.: Negative human rights as a basis for long-term AI safety and regulation. J. Artif. Intell. Res. **76**, 1043–1075 (2023). https://doi.org/10.1613/jair.1.14020
8. Bächle, T. C., & Bareis, J. (eds.).: The Realities of Autonomous Weapons. Bristol University Press (2025)
9. Bareis, J.: Ask Me Anything! How ChatGPT Got Hyped Into Being. OSF (2024a). https://doi.org/10.31235/osf.io/jzde2
10. Bareis, J.: The trustification of AI. Disclosing the bridging pillars that tie trust and AI together. Big Data Soc. (2024). https://doi.org/10.1177/20539517241249430
11. Bareis, J., Roßmann, M., & Bordignon, F.: Technology hype: Dealing with bold expectations and overpromising. TATuP - Zeitschrift Für Technikfolgenabschätzung in Theorie und Praxis, 32(3) (2024b). https://doi.org/10.14512/tatup.32.3.10
12. Bauer-Kahan, R.: ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION (Hearing SB 1047). https://apcp.assembly.ca.gov/system/files/2024-06/sb-1047-wiener-apcp-analysis_0.pdf (2024). Accessed 13 August 2025
13. Beckert, J.: Imagined Futures: Fictional Expectations and Capitalist Dynamics. Harvard University Press (2016)
14. Bender, E. M., & Hanna, A.: AI Causes Real Harm. Let's Focus on That over the End-of-Humanity Hype. Scientific American. https://www.scientificamerican.com/article/we-need-to-focus-on-ais-real-harms-not-imaginary-existential-risks/ (2023). Accessed 13 August 2025
15. Benjamin, R.: Imagination: A manifesto. WW Norton & Company (2024)
16. Beltramini, E.: The government of evil machines: an application of Romano Guardini's thought on technology. Sci. Fides **9**(1), 275–281 (2021). https://doi.org/10.12775/setf.2021.010
17. Blili-Hamelin, B., Hancox-Li, L., & Smart, A.: Unsocial Intelligence: An Investigation of the Assumptions of AGI Discourse. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7(1), 141-155 (2024). https://doi.org/10.1609/aies.v7i1.31625
18. Blili-Hamelin, B., Graziul, C., Hancox-Li, L., Hazan, H., El-Mhamdi, E.-M., Ghosh, A., Heller, K., Metcalf, J., Murai, F., Salvaggio, E., Smart, A., Snider, T., Tighanimine, M., Ringer, T., Mitchell, M., Dori-Hacohen, S.: Stop treating `AGI' as the north-star goal of AI research. arXiv (2025). https://doi.org/10.48550/arXiv.2502.03689
19. Bostrom, N.: Existential risk prevention as global priority. Glob. Policy **4**(1), 15–31 (2013)
20. Brock, D. C.: Our Censors, Ourselves: Commercial Content Moderation. Los Angeles Review of Books. https://lareviewofbooks.org/article/our-censors-ourselves-commercial-content-moderation (2019). Accessed 13 August 2025
21. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, Ma.K., Schwitzgebel, E., Simon, J., VanRullen, R.: Consciousness in artificial intelligence: insights from the science of consciousness. arXiv (2023). https://doi.org/10.48550/arXiv.2308.08708
22. Campolo, A., Crawford, K.: Enchanted determinism: power without responsibility in artificial intelligence. Engag. Sci. Technol. Soc. (2020). https://doi.org/10.17351/ests2020.277
23. Cave, S., Dihal, K.: Hopes and fears for intelligent machines in fiction and reality. Nat. Mach. Intell. **1**(2), 74–78 (2019). https://doi.org/10.1038/s42256-019-0020-9
24. Center for AI Safety: Statement on AI Risk - AI experts and public figures express their concern about AI risk. https://www.safe.ai/work/statement-on-ai-risk (2023). Accessed 13 August 2025
25. Coeckelbergh, M.: Can we trust robots? Ethics Inf. Technol. **14**(1), 53–60 (2012). https://doi.org/10.1007/s10676-011-9279-1
26. Courtial, J.-P., Law, J.: A co-word study of artificial intelligence. Soc. Stud. Sci. **19**(2), 301–311 (1989). https://doi.org/10.1177/030631289019002005
27. Cremer, C.Z., Whittlestone, J.: Artificial canaries: early warning signs for anticipatory and democratic governance of AI. Int. J. Interact. Multimed. Artif. Intell. **6**(5), 100 (2021). https://doi.org/10.9781/ijimai.2021.02.011
28. Cronin, M.A., George, E.: The why and how of the integrative review. Organ. Res. Methods **26**(1), 168–192 (2023). https://doi.org/10.1177/1094428120935507
29. Depaz, P.: Shaping vectors: discipline and control in word embeddings. A Peer-Rev. J. About **13**(1), 90–104 (2024). https://doi.org/10.7146/aprja.v13i1.151234
30. Dung, L.: How to deal with risks of AI suffering. Inquiry (2023). https://doi.org/10.1080/0020174X.2023.2238287
31. Dung, L.: Current cases of AI misalignment and their implications for future risks. Synthese (2023). https://doi.org/10.1007/s11229-023-04367-0
32. Dung, L.: The argument for near-term human disempowerment through AI. AI & Soc. (2024). https://doi.org/10.1007/s00146-024-01930-2
33. Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., & Fernandez-Llorca, D.: Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. arXiv preprint (2025). https://doi.org/10.48550/arXiv.2502.06559
34. Evers, K., Farisco, M., Chatila, R., Earp, B.D., Freire, I.T., Hamker, F., Németh, E., Verschure, P.F.M.J., Khamassi, M.: Preliminaries to artificial consciousness: a multidimensional heuristic approach. Phys. Life Rev. (2025). https://doi.org/10.1016/j.plrev.2025.01.002
35. Frey, P., Dobroć, P., Hausstein, A., Heil, R., Lösch, A., Roßmann, M., Schneider, C.: Vision Assessment: Theoretische Reflexionen zur Erforschung soziotechnischer Zukünfte. KIT Sci. Publish. (2022). https://doi.org/10.5445/KSP/1000142150
36. Future of Life Institute: Pause Giant AI Experiments: An Open Letter. Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (2023). Accessed 13 August 2025
37. Gebru, T., Torres, É.P.: The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. First Monday **29**, 4 (2024). https://doi.org/10.5210/fm.v29i4.13636
38. Gibbs, S.: Elon Musk: Artificial intelligence is our biggest existential threat. The Guardian. https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat (2014). Accessed 13 August 2025

39. Goffman, E.: The Presentation of Self in Everyday Life. In W. Longhofer, & D. Winchester (eds.) Social Theory Re-Wired. New Connections to Classical and Contemporary Perspectives (3rd Ed.). Routledge (2023)
40. Goldstein, S., Kirk-Giannini, C.D.: Language agents reduce the risk of existential catastrophe. AI Soc. **40**, 959–969 (2025). https://doi.org/10.1007/s00146-023-01748-4
41. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O.: Viewpoint: when will AI exceed human performance? Evidence from AI experts. J. Artif. Intell. Res. **62**, 729–754 (2018). https://doi.org/10.1613/jair.1.11222
42. Graham, R.: Discourse analysis of academic debate of ethics for AGI. AI & Soc. **37**(4), 1519–1532 (2022). https://doi.org/10.1007/s00146-021-01228-7
43. Grin, J.: Vision Assessment to Support Shaping 21st Century Society? Technology Assessment as a Tool for Political Judgement. In: Grin, J., Grunwald, A. (eds) Vision Assessment: Shaping Technology in 21st Century Society. Wissenschaftsethik und Technikfolgenbeurteilung, vol 4. Springer, Berlin, Heidelberg (2000). https://doi.org/10.1007/978-3-642-59702-2_2
44. Grunwald, A.: The hermeneutic side of responsible research and innovation. J. Responsib. Innov. **1**(3), 274–291 (2014). https://doi.org/10.1080/23299460.2014.968437
45. Grunwald, A.: Die hermeneutische Erweiterung der Technikfolgenabschätzung. TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis (2015). https://doi.org/10.14512/tatup.24.2.65
46. Gyevnar, B., Kasirzadeh, A.: AI safety for everyone. Nat. Mach. Intell. (2025). https://doi.org/10.1038/s42256-025-01020-y
47. Hadshar, R.: A review of the evidence for existential risk from AI via misaligned power-seeking. arXiv (2023). https://doi.org/10.48550/arXiv.2310.18244
48. Hanson, R., Yudkowsky, E.: AI-Foom Debate. Machine Intelligence Research Institute (2013).
49. Harvey, I.: Motivations for Artificial Intelligence, for Deep Learning, for ALife: mortality and existential risk. Artif. Life **30**(1), 48–64 (2024). https://doi.org/10.1162/artl_a_00427
50. Heil, J.: Philosophy of Mind: A Contemporary Introduction (4th ed.). Routledge (2019). https://doi.org/10.4324/9780429506994
51. Hyde, B.V.E.: The problem with longtermism. Ethics Prog. **14**(2), 130–152 (2023). https://doi.org/10.14746/eip.2023.2.9
52. Jebari, K., Lundborg, J.: Artificial superintelligence and its limits: why AlphaZero cannot become a general agent. AI & Soc. **36**, 807–815 (2021). https://doi.org/10.1007/s00146-020-01070-3
53. Jilk, D.J.: Conceptual-linguistic superintelligence. Informatica **41**(4), 429–439 (2017)
54. Johnson, D.G., Verdicchio, M.: Reframing AI discourse. Minds Mach. **27**(4), 575–590 (2017). https://doi.org/10.1007/s11023-017-9417-6
55. Kapoor, S., & Narayanan, A.: A misleading open letter about sci-fi AI dangers ignores the real risks. Ai Snake Oil https://www.aisnakeoil.com/p/a-misleading-open-letter-about-sci (2023). Accessed 13 August 2025
56. Konrad, K. E., Lente, H. van, Groves, C., & Selin, C.: Performing and Governing the Future in Science and Technology. In U. Felt, R. Fouche, C. A. Miller, & L. Smith-Doerr (eds.) The Handbook of Science and Technology Studies (4th Edition), pp. 465–493. MIT Press (2016).
57. Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd Edition, enlarged). Univ. of Chicago Press. https://www.lri.fr/~mbl/Stanford/CS477/papers/Kuhn-SSR-2ndEd.pdf
58. Lavazza, A., Vilaça, M.: Human extinction and AI: what we can learn from the ultimate threat. Philos. Technol. (2024). https://doi.org/10.1007/s13347-024-00706-2
59. Leggett, D.: Feeding the Beast: Superintelligence, Corporate Capitalism and the End of Humanity. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society 727–735 (2021). https://doi.org/10.1145/3461702.3462581
60. Mackenzie, A.: The production of prediction: what does machine learning want? Eur. J. Cult. Stud. **18**(4–5), 429–445 (2015). https://doi.org/10.1177/1367549415577384
61. Makridakis, S., Bakas, N.: Forecasting and uncertainty: a survey. Risk Decis. Anal. **6**(1), 37–64 (2016). https://doi.org/10.3233/RDA-150114
62. Martino, J.P.: A review of selected recent advances in technological forecasting. Technol. Forecast. Soc. Change **70**(8), 719–733 (2003). https://doi.org/10.1016/S0040-1625(02)00375-X
63. Maskara, P.K.: Developing safer AI–concepts from economics to the rescue. AI & Soc. **40**, 971–983 (2023). https://doi.org/10.1007/s00146-023-01778-y
64. McGoey, L.: The Unknowers: How Strategic Ignorance Rules the World. Bloomsbury Publishing (2019).
65. McLean, S., Read, G.J.M., Thompson, J., Baber, C., Stanton, N.A., Salmon, P.M.: The risks associated with Artificial General Intelligence: a systematic review. J. Exp. Theor. Artif. Intell. **35**(5), 649–663 (2023). https://doi.org/10.1080/0952813X.2021.1964003
66. Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T.: Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review. 2016 Portland International Conference on Management of Engineering and Technology (PICMET), 682–693 (2016). https://doi.org/10.1109/PICMET.2016.7806752
67. Merchant, B.: In California, no AI bill is safe. Blood in the Machine https://www.bloodinthemachine.com/p/in-california-no-ai-bill-is-safe (2024). Accessed 13 August 2025
68. Minsky, M. L.: Computation: Finite and infinite machines. Englewood Cliffs, N.J. Prentice-Hall (1967).
69. Mitchell, M.: Debates on the nature of artificial general intelligence. Science (2024). https://doi.org/10.1126/science.ado7069
70. Mosco, V.: The Digital Sublime: Myth, Power, and Cyberspace. The MIT Press (2004)
71. Natale, S., Ballatore, A.: Imagining the thinking machine: technological myths and the rise of artificial intelligence. Convergence **26**(1), 3–18 (2020). https://doi.org/10.1177/1354856517715164
72. Nathan, C., Hyams, K.: Global catastrophic risk and the drivers of scientist attitudes towards policy. Sci. Eng. Ethics (2022). https://doi.org/10.1007/s11948-022-00411-3
73. Nordmann, A.: If and Then: A Critique of Speculative NanoEthics. In A. Maynard, & J. Stilgoe (eds.) The Ethics of Nanotechnology, Geoengineering, and Clean Energy, pp. 31–46. Routledge (2017)
74. Oomen, J., Hoffman, J., Hajer, M.A.: Techniques of futuring: on how imagined futures become socially performative. Eur. J. Soc. Theory **25**(2), 252–270 (2022). https://doi.org/10.1177/1368431020988826
75. Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C.B.C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., Hendrycks, D.: Humanity's last exam. arXiv (2025). https://doi.org/10.48550/arXiv.2501.14249
76. Perry, B., Uuk, R.: AI governance and the policymaking process: key considerations for reducing AI risk. Big Data Cogn. Comput. (2019). https://doi.org/10.3390/bdcc3020026
77. Porter, T.: Trust in numbers: the pursuit of objectivity in science and public life. Princeton University Press, Princeton (1995)
78. Raji, I.D., Bender, E.M., Paullada, A., Denton, E., Hanna, A.: AI and the everything in the whole wide world benchmark. arXiv (2021). https://doi.org/10.48550/arXiv.2111.15366
79. Renn, O.: Risk governance: coping with uncertainty in a complex world. Routledge (2008). https://doi.org/10.4324/9781849772440

80. Roose, K.: When A.I. Passes This Test, Look Out. The New York Times. https://www.nytimes.com/2025/01/23/technology/ai-test-humanitys-last-exam.html (2025). Accessed 13 August 2025
81. Ryazanov, I., Öhman, C., Björklund, J.: How ChatGPT changed the media's narratives on AI: a semi-automated narrative analysis through frame semantics. Minds Mach. **35**(1), 2 (2025). https://doi.org/10.1007/s11023-024-09705-w
82. Saavedra-Rivano, N.: AI and Us: Existential Risk or Transformational Tool? In: Proceedings of 2019 IEEE 18th International Conference on Cognitive Informatics and Cognitive Computing, pp. 319–322 (2019)
83. Shestakofsky, B.: Behind the Startup: How venture capital shapes work, innovation, and inequality. University of California Press (2024)
84. Shiller, D.: In defense of artificial replacement. Bioethics **31**(5), 393–399 (2017). https://doi.org/10.1111/bioe.12340
85. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T.: Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint (2017). https://doi.org/10.48550/arXiv.1712.01815
86. Simon, H.A.: The new science of management decision. Harper & Brothers (1960). https://doi.org/10.1037/13978-000
87. Singler, B.: Existential hope an existential despair in AI. Zygon **54**(1), 156–176 (2019). https://doi.org/10.1111/zygo.12494
88. Sotala, K., Yampolskiy, R.V.: Responses to catastrophic AGI risk: a survey. Phys. Scr. (2015). https://doi.org/10.1088/0031-8949/90/1/018001
89. Suchman, L., & Weber, J.: Human–machine autonomies. In C. Kreβ, H.-Y. Liu, N. Bhuta, R. Geiβ, & S. Beck (eds.) Autonomous Weapons Systems: Law, Ethics, Policy, pp. 75–102. Cambridge University Press (2016). https://doi.org/10.1017/CBO9781316597873.004
90. Suchman, L.: The uncontroversial 'thingness' of AI. Big Data Soc. (2023). https://doi.org/10.1177/20539517231206794
91. Thorstad, D.: Against the singularity hypothesis. Philos. Stud. **182**, 1627–1651 (2024). https://doi.org/10.1007/s11098-024-02143-5
92. Torraco, R.J.: Writing integrative literature reviews: guidelines and examples. Hum. Resour. Dev. Rev. **4**(3), 356–367 (2005). https://doi.org/10.1177/1534484305278283
93. Torres, P.: The possibility and risks of artificial general intelligence. Bull. At. Sci. **75**(3), 105–108 (2019). https://doi.org/10.1080/00963402.2019.1604873
94. Torres, É. P.: The Dangerous Ideas of Longtermism and "Existential Risk". Current Affairs https://www.currentaffairs.org/news/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk (2021). Accessed 13 August 2025
95. Torres, É. P.: Longtermism poses a real threat to humanity. New Statesman. https://www.newstatesman.com/ideas/2023/08/longtermism-threat-humanity (2023). Accessed 13 August 2025
96. Turchin, A.: Assessing the future plausibility of catastrophically dangerous AI. Futures **107**, 45–58 (2019). https://doi.org/10.1016/j.futures.2018.11.007
97. Turchin, A., Denkenberger, D.: Classification of global catastrophic risks connected with artificial intelligence. AI & Soc. **35**(1), 147–163 (2018). https://doi.org/10.1007/s00146-018-0845-5
98. Undheim, T.A.: An interdisciplinary review of systemic risk factors leading up to existential risks. Prog. Disaster Sci. **22**, 100326 (2024). https://doi.org/10.1016/j.pdisas.2024.100326
99. van der Vlist, F., Helmond, A., Ferrari, F.: Big AI: cloud infrastructure dependence and the industrialisation of artificial intelligence. Big Data Soc. (2024). https://doi.org/10.1177/20539517241232630
100. von der Leyen, U.: State of the Union Address by President von der Leyen. European Commission. https://ec.europa.eu/commission/presscorner/detail/ov/speech_23_4426 (2023). Accessed 13 August 2025
101. Wallach, W., Allen, C.: Moral machines: teaching robots right from wrong. Oxford University Press (2008). https://doi.org/10.1093/acprof:oso/9780195374049.001.0001
102. Wenar, L.: The Deaths of Effective Altruism. Wired. https://www.wired.com/story/deaths-of-effective-altruism/ (2024). Accessed 13 August 2025
103. Wiebe, K., Zurek, M., Lord, S., Brzezina, N., Gabrielyan, G., Libertini, J., Loch, A., Thapa-Parajuli, R., Vervoort, J., Westhoek, H.: Scenario development and foresight analysis: exploring options to inform choices. Annu. Rev. Environ. Resour. **43**, 545–570 (2018). https://doi.org/10.1146/annurev-environ-102017-030109
104. Winner, L.: Autonomous technology: Technics-out-of-control as a theme in political thought. Mit Press (1978).
105. Wold, J. W.: Academics to chair drafting the Code of Practice for general-purpose AI. Euractiv. https://www.euractiv.com/section/tech/news/academics-to-chair-drafting-the-code-of-practice-for-general-purpose-ai/ (2024). Accessed 13 August 2025
106. Woolgar, S.: Why not a sociology of machines? The case of sociology and artificial intelligence. Sociology **19**(4), 557–572 (1985). https://doi.org/10.1177/0038038585019004005
107. Yamakawa, H.: Peacekeeping conditions for an artificial intelligence society. Big Data Cogn. Comput. **3**(2), 34 (2019). https://doi.org/10.3390/bdcc3020034
108. Yampolskiy, R.V.: AGI Control Theory. In: Goertzel, B., Iklé, M., Potapov, A. (eds) Artificial General Intelligence. AGI 2021. Lecture Notes in Computer Science, vol 13154. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-93758-4_33
109. Yudkowsky, E.: There's No Fire Alarm for Artificial General Intelligence. Machine Intelligence Research Institute. https://intelligence.org/2017/10/13/fire-alarm/ (2017). Accessed 13 August 2025

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