AI and Ethics (2025) 5:2227–2242
https://doi.org/10.1007/s43681-024-00548-w

ORIGINAL RESEARCH

On singularity and the Stoics: why Stoicism offers a valuable approach to navigating the risks of AI (Artificial Intelligence)

Bernardo Bolaños Guerra[1] · **Jorge Luis Morton Gutierrez**[1]

Received: 22 April 2024 / Accepted: 4 August 2024 / Published online: 14 August 2024

Abstract
The potential benefits and risks of artificial intelligence technologies have sparked a wide-ranging debate in both academic and public circles. On one hand, there is an urgent call to address the immediate and avoidable challenges associated with these tools, such as accountability, privacy, bias, understandability, and transparency; on the other hand, prominent figures like Geoffrey Hinton and Elon Musk have voiced concerns over the potential rise of Super Artificial Intelligence, whose singularity could pose an existential threat to humanity. Coordinating the efforts of thousands of decentralized entities to prevent such a hypothetical event may seem insurmountable in our intricate and multipolar world. Thus, drawing from both perspectives, this work suggests employing the tools and framework of Stoic philosophy, particularly the concept of the dichotomy of control—focusing on what is within our power. This Stoic principle offers a practical and epistemological approach to managing the complexities of AI, and it encourages individuals to organize their efforts around what they can influence while adapting to the constraints of external factors. Within this framework, the essay found that Stoic wisdom is essential for assessing risks, courage is necessary to face contemporary challenges, and temperance and tranquility are indispensable; and these lessons can inform ongoing public and academic discourse, aiding in the development of more effective policy proposals for aligning Narrow AI and General AI with human values.

Keywords Super AI · Generative AI · Risk · Control · Narrow AI · Singularity

1 Introduction

Artificial intelligence (AI) is one of the most transformative and debated technologies of our time, with the potential to revolutionize aspects of human life such as healthcare, education, communication, and entertainment. However, AI also poses significant risks and challenges in areas like human rights, social welfare, and global security. The pivotal question is how to guide the responsible and ethical development and utilization of AI in a way that benefits humanity rather than causing harm, especially considering the current and potential risks associated with the technology and the possibility of a Super Artificial Intelligence (Super AI) emerging.

Within this context, it is crucial to recognize that AI is a significant technology that spans various scientific and engineering disciplines, enabling machines to learn by "memorizing particular facts or brand-new information" [3]. While there is no universally agreed-upon definition of intelligence, "but one aspect that is broadly accepted is that intelligence is not limited to a specific domain or task, but rather encompasses a broad range of cognitive skills and abilities" [8]. Current AI technologies exhibit a wide array of capabilities, from enhancing daily experiences like providing optimal navigation routes to performing complex tasks such as generating textual content based on user inputs; thus, a broad definition of Artificial Intelligence could be "the study, design, and building of intelligent agents that can achieve goals" [6]; AI can be categorized into Narrow AI, which is "goal-oriented and performs specific tasks" [3], and General AI, which, though it does not yet exist, refers to technologies with "the capability to understand and think similarly to humans." (ibid).

Besides defining AI, understanding and refining the taxonomies of AI is important for creating a regulatory

✉ Bernardo Bolaños Guerra
bbolanos@cua.uam.mx

Jorge Luis Morton Gutierrez
jorge.morton@cua.uam.mx

[1] Universidad Autónoma Metropolitana, Cuajimalpa, Mexico City, Mexico

framework. Revising the risk taxonomies is crucial because it can anchor public policy development related to AI regulation. Also, it is essential to acknowledge the anxiety surrounding the potential emergence of a technology where "technological growth becomes uncontrollable and irreversible, resulting in unforeseeable changes to human civilization" [28].

Western media has often depicted AI as a technology that could potentially cause significant harm to humanity. While some narratives present AI in a non-threatening light, the predominant stories emphasize catastrophic outcomes that could arise from its misapplication [24] Historically, this perspective was largely confined to pop culture, but in recent years, prominent public figures like Elon Musk and AI research organizations such as OpenAI have highlighted the risks associated with the development of Artificial General Intelligence (AGI) and the subsequent possibility of a Super AI; moreover it is crucial to recognize that not all AI poses the same level of risk, and grounding policy discussions within this context is essential for effective regulation and public understanding.

The European Union's AI Act serves as a key example, adopting a risk-based framework to classify AI technologies and shape regulatory measures [1]. This legislation emphasizes the need for a solid theoretical and analytical foundation, essential for crafting comprehensive policies that harmonize innovation with ethical considerations and the public good [15]. However, the regulation does not fully address the ongoing debate about the risks of Narrow AI, AGI, or Super AI.

The notion of the dichotomy of control in Stoic philosophy offers a referential framework for assessing the avoidable, current, immediate, and potential risks of AI. It proposes a more balanced approach to addressing these risks while capitalizing on the technology's societal benefits. Within this context, the current essay will unfold as follows: Sect. 2 introduces pertinent notions of Stoic philosophy and its relevance to this essay, alongside the debate that sparked the elaboration of this work. Section 3 examines the taxonomies of AI risk. Section 4 focuses on discussions around the fears of a Super AI, their impact on public policy and the research and development of this technology. Section 5 explores how Stoic philosophy can aid in cultivating resilience, wisdom, and ethical decision-making in the face of AI challenges. Finally, Sect. 6 concludes with the scope of this study and areas of opportunity for future research.

2 On Stoic philosophy

This essay delves into Stoic philosophy, an ancient Greek school of thought established by Zeno of Citium in the early third century BCE. Stoic philosophy underscores the importance of concentrating on what is within our control and accepting what is not, a principle known as the "dichotomy of control" ([43], pp. 31–27). And several renowned Stoic philosophers—notably Chrysippus, Marcus Aurelius, Epictetus, and Seneca—emphasized this concept; for example, Marcus Aurelius, a Roman Emperor, reflected on this in his writings, the *Meditations*, noting that internal thoughts and attitudes are within our control, whereas external events are not by stating "…things do not take hold upon the mind but stand without unmoved, and that disturbances come only from the judgment within" (*Meditations*, IV.3).

The Stoics believed in fate as an "ordering and sequence of causes" that determines every fact in the past, present, and future (Cicero, *On Divination*, 1.125-6) [11]. Chrysippus (c. 279–c. 206 BC) distinguished between simple and conjoined outcomes. Simple outcomes, such as the death of a mortal, occur regardless of other events. Conjoined outcomes, however, occur as linked cause and effect: "For instance, if Oedipus is going to be born of Laius, that is conjoined with Laius and Jocasta having intercourse" [41]. In the case of the fated emergence of a Super AI aligned with human values, that alignment would be conjoined with the actions of investing sufficient resources to develop responsible technology.

Epictetus (c. 50–c. 135 AD), a Greek Stoic philosopher and former slave, famously articulated the Stoic doctrine of control in his *Enchiridion*, stating: "Some things are within our power, while others are not. Within our power are opinion, motivation, desire, aversion, and, in a word, whatever is of our own doing; not within our power are our body, our property, reputation, office, and, in a word, whatever is not of our own doing" (Enchiridion, I.1). This principle emphasizes focusing on what we can control and accepting what we cannot. Seneca the Younger (c. 4 BC–AD 65) echoed these sentiments in his writings. In Letter XIII.14, he wrote: "Let another say: 'Perhaps the worst will not happen'. You yourself must say: 'Well, what if it does happen? Let us see who wins! Perhaps it happens for my best interests'". This reflects the Stoic attitude of preparedness and resilience in the face of adversity.

Thus, Stoicism teaches that our character, whether we are virtuous or knowledgeable, is in our control, while some external events are not. Marcus Aurelius (121–180 AD) advised:

> "What shall we say? Wait in peace, whether for extinction or a change of state; and until its due time arrives, what is sufficient? What else than to worship and bless the gods, to do good to men, to bear them and to forbear; and, for all that lies within the limits of mere flesh and spirit, to remember that this is neither yours nor in your power" (*Meditations*, V.33).

Epictetus also remarked, "When I see someone in anxiety, I say to myself, what can it be that this fellow wants?

For if he did not want something that was outside of his control, how could he remain in anxiety?" (*Discourses*, Book II, ch. 13, § 1) [14]. In this sense, Stoic philosophy serves as a practical tool to address various issues surrounding AI. This field has seen a resurgence, thanks to authors like A. A. Long, Gisela Striker, Julia Annas, Pierre Hadot, Martha Nussbaum, Brad Inwood, John Sellars, and Laurence Becker [52]. What makes Stoicism a valuable perspective for engaging with AI risk discussions? To answer this Spence [52] comments:

> Stoicism encourages active political and social engagement and not merely a retreat to a communal garden; unlike Cynicism, Stoicism considers preferred indifferents like health, wealth and social status as desirable so long as they accord with a virtuous lifestyle. The Stoics, however, concur with the Cynics that only virtue is good and that it alone is both necessary and sufficient for a good and happy life, a eudaimonic life.

In addition, Spence [52] outlines seven key principles of Stoic and Neo-Stoic Philosophy. First, the concern should be directed only towards things within our control; Second, the pursuit of *eudaimonia*, or well-being, is essential; Third, one should strive to live a life characterized by virtue and wisdom; Fourth, cosmopolitanism advocates for collective action; Fifth, living a life in agreement with nature is emphasized; Sixth, *oikeiosis* is highlighted as the practical necessity for moral transformation, offering a framework for such change in practice. Lastly, philosophy is presented as a way of life. All these principles contribute to the broader discourse on AI and ethics; however, the discussion of the dangers of Narrow AI and AGI will focus primarily on the principle of control.

Stoic philosophy teaches that we have control over our mental and behavioral responses, such as our opinions, desires, and personal choices. In contrast, we do not control the behavior of others, world events, or natural occurrences. As Marcus Aurelius, Epictetus, and Seneca, among other Stoics, instruct, we should direct our efforts towards areas within our influence and maintain a serene acceptance of those beyond our control.

However, Stoicism does not promote passivity or indifference in adversity. Instead, it encourages individuals to face challenges with rationality, temperance, tranquility, and virtue or excellence in general (*arete*), acknowledging that while we cannot always control external events, we can control our reactions to them. As Epictetus stated: "Come now, haven't you been endowed with faculties that enable you to bear whatever may come about? Haven't you been endowed with greatness of soul? And with courage? And with endurance?" (*Discourses,* Book I, ch. 6, § 28) [14].

Stoicism does not suggest that we ignore the immutable; rather, it teaches us to acknowledge and accept what cannot be changed while focusing our efforts on what we can control. The philosophy encourages a balanced approach, recognizing the existence of unchangeable external factors but emphasizing the importance of our internal responses and attitudes toward them. Therefore, Stoic philosophy advocates facing challenges with courage, prudence, and resilience. This tenet is reinforced by the doctrines of eminent Stoic philosophers such as Chrysippus, Marcus Aurelius, Epictetus, and Seneca, who all emphasized the "dichotomy of control," differentiating between what falls within our power and what does not.

Marcus Aurelius famously stated that true power lies in our minds, not in external circumstances, and recognizing this is a source of strength (Aurelius, *Meditations*, Book IV, Sect. 5): lessons that were valuable for him in his duties as both a general and an emperor. Seneca follow these sentiments in his works, particularly in "On the Shortness of Life", where he discusses the importance of concentrating on what is within our control, aligning with the Stoic focus on inner virtues and choices by stating:

> "We are in the habit of saying that it was not in our power to choose the parents who were allotted to us, that they were given to us by chance. But we can choose whose children we would like to be. There are households of the noblest intellects: choose the one into which you wish to be adopted, and you will inherit not only their name but their property too" (15.3) [49].

While the Stoics encourage us to align ourselves with the teachings and virtues of prudent individuals rather than irresponsible alarmists, they also advocate for open-mindedness, reminding us of the importance of listening more than speaking, as symbolized by our having one mouth but two ears. These esteemed Stoic philosophers collectively emphasize focusing on what lies within our realm of influence—our thoughts, beliefs, attitudes, and actions—while accepting with equanimity those aspects beyond our control, such as external events, the actions of others, and natural occurrences. And this philosophical stance enables us to tackle the pressing issues in the debate over the risks of Narrow AI, AGI, and potentially Super AI; however, it is imperative first to examine the division that has sparked the disputes fracturing many connections within this discourse.

2.1 On the AI schism

The emergence of Generative AI (Gen AI) models like Claude-3 and ChatGPT has sparked a significant debate within the AI community. Before the public unveiling of these models, discussions on AI risks and regulations primarily focused on the acknowledged dangers of Narrow AI.

Authors across various fields, such as O'Neill [42] and Van Dijck et al. [54], have extensively documented these risks, which include concerns about privacy, discrimination, and information access. These issues are particularly pertinent in areas where machine learning models are applied, such as surveillance with facial recognition technology, healthcare, commerce, and media and platforms like Netflix.

Generative AI models have introduced increased risk and complexity in regulation. Many researchers and experts commend ChatGPT for its advancements in natural language processing and its potential for diverse applications. However, there are also voices raising ethical and societal concerns, pointing out the dangers of spreading misinformation, manipulation, bias, and deception. Consequently, the conversation surrounding Generative AI has expanded to encompass debates on the limitations and responsibilities tied to AI's development and use.

Sam Altman, CEO of OpenAI—the organization behind ChatGPT—has cautioned against the risks of misinformation and "hallucination" (the fabrication of information) that these models may present. Altman advises, "The appropriate way to perceive the models we create is as reasoning engines, not repositories of facts" [21]. While ChatGPT can generate coherent and logical text, it may not always be factual or accurate and users are encouraged to approach interactions with ChatGPT with caution and critical thinking, rather than treating it as a reliable source of information.

Altman presented his concerns to the US Congress [77], discussing the potential risks of ChatGPT and Large Language Models (LLMs). Nonetheless, several researchers and engineers argue that the fear of malevolent AI distracts from the actual, current problems caused by technology. These problems extend beyond human rights issues to include a lack of accountability and fairness in AI systems, as well as their impact on security, democratic processes, and human dignity.

OpenAI has come under scrutiny for not disclosing the training methods of ChatGPT, raising concerns about the human right to information access. Sam Altman, criticizing the EU AI Act, remarked, "The current draft of the EU AI Act would be overly restrictive, but we've heard it's going to be revised" [62]. And he suggested that such restrictions could hinder AI innovation, which has potential applications in addressing global challenges like climate change.

Considering the capabilities of ChatGPT and other large language models (LLMs), they serve as valuable tools in research and academia, assisting with grammatical corrections and basic editing. For example, the authors of this paper utilized ChatGPT for grammar refinement. Additionally, LLMs like Chat GPT can tentatively help "to grade and provide feedback on student assignments" [32]. However, these models also present risks in education, such as enabling students to "cheat on essay writing assignments by feeding the chatbot specific prompts and questions, and then copying and pasting the generated responses into their own papers" [26]. Moreover, text-to-image models pose additional challenges to academic integrity by sometimes bypassing peer review processes with misleading outputs [30]; and the ever-present risk of hallucinations (making things up) suggest that people should be taught about the limitations and capabilities of these models with practices such as prompt engineering (using the best instructions) that can help improve their work in areas such as proofreading, ensuring that the output is both accurate and reliable.

Now, the debate on whether large language models (LLMs) utterly understand central to regulatory discussions, requiring insights from contemporary philosophy and data science. This discourse, as emphasized by Hinton (initially) and Ng [68], should focus on the immediate challenges posed by these technologies. Additionally, we can draw from Stoicism, which distinguishes between avoidable and unavoidable risks, to inform our understanding of AI. Claims of self-awareness in models like Claude-3 [71] have intensified debates on the actual, present risks of AI, contrasting with the speculative fears of a singularity event.

The debate on Narrow AI versus AGI and Super AI is particularly pressing, given the division it has caused among the public, media, and academic community. Some view the Super Intelligence risk debate as a diversion by corporations from the urgent issues current AI models present. For instance, Geoffrey Hinton has been caught up in this controversy. Timnit Gebru expressed skepticism at Rights Con, questioning the sudden shift in narrative around GPT-4: "Just a few months ago, Geoff Hinton was discussing GPT-4 and likening it to a caterpillar that consumes data and then transforms into a magnificent butterfly; now, suddenly, it is being portrayed as an existential risk. I mean, why are people taking these individuals seriously?" (Ryan-Mosley, 2023).

The perceived level of risk differs from the existential risks (X-risks) described by Karina Vold and Daniel R. Harris [56] as the most significant threats to humanity. This raises the question: should we categorize current AI as a risk based on misconceptions about its capabilities? Given the current technological state, the answer might be no. However, it is crucial to address the risks associated with existing Narrow AI models, even if they do not pose a catastrophic threat. Just as we cannot afford to ignore the consequences of climate change and the need for preventative measures, we must also proactively address the immediate and foreseeable risks of AI. This involves cutting through the noise and distractions from various camps—those focusing on immediate risks and those undermining current problems by shifting attention to non-existent scenarios. It is essential to address both the immediate concerns and the potential for

the most catastrophic scenarios that general artificial intelligence might pose.

Now, the division between high-risk and immediate-risk perceptions of AI may stem from several factors:

1. The significant attention garnered by the LLM model upon its public release in 2022. The public and academia outside of AI research were introduced to a new kind of generative AI capable of not just recommending movies on Netflix but also predicting and generating text.
2. The scrutiny these models face within regulatory frameworks, such as the AI Act in Europe [16, 17].
3. Public statements by figures like Elon Musk, who have voiced concerns about AI risks.
4. The ability of LLMs to emulate and reproduce human language—a quality many philosophers and humanists value—through computational means, contributing to a narrative that diverts attention from the technology's immediate and avoidable problems and pressing issues.

Finally, the replication of human language by LLMs through computational power, algorithm technology and bast amounts of data, has led to a narrative that overshadows the immediate and pressing issues of the technology, as demonstrated by the attention given to ChatGPT and the speculative concept of Super AI. This has resulted in a schism and considerable noise regarding the problems of narrow and general AI, which can be better understood through a Stoic perspective of control. However, for narrow AI, there are taxonomies that explain the associated risks, providing a foundation to address policy and regulation effectively.

3 On taxonomies of risk of AI

To comprehend the challenges posed by current AI models, it's crucial to categorize the risks they present. This section draws upon the risk taxonomies established by the National Institute of Standards and Technology (NIST) and Newman's work on a taxonomy of trustworthiness for artificial intelligence, which connects trustworthiness attributes with risk management and the AI lifecycle (Center for Long-Term Cybersecurity, 2023). These frameworks address risks related to technical and design features, perceptions of AI systems, regulatory policies and principles, as well as environmental impacts.

NIST [40] identifies four key elements to evaluate technical risks: Accuracy, Reliability, Robustness, and Resilience or Security. Accuracy is the AI system's ability to yield correct and consistent results. Reliability refers to its performance under normal conditions. Robustness denotes the system's capacity to cope with adverse or unforeseen situations. Resilience or Security involves the system's ability to recover from disruptions or damage. Blackman [5] distinguishes between two AI development approaches: "AI for Good," which seeks positive societal impact, and "AI for not Bad," which aims to avoid ethical missteps while pursuing various objectives, whether ethically commendable or neutral.

Regarding socio-technical attributes, NIST [40] considers how AI-related risks are measured, the systems' broader implications for users, and the influence on human judgment. These attributes include explicability, which clarifies model predictions, and interpretability, which effectively contextualizes model outputs. NIST also emphasizes the importance of addressing privacy, safety, and bias in AI models and systems.

Blackman [5] points out that bias, privacy, and explainability are the three most significant AI risks. At forums on AI ethics, discussions often center on AI bias, the opacity of algorithms, and privacy violations. Models trained on biased data risk perpetuating those biases and pose significant privacy threats concerning the data they're trained on and consumer data. Additionally, the models act as "black boxes" because it's challenging for users to understand how predictions are made, or the operational details are kept confidential by developers and companies.

The significance of interpretability in AI cannot be overstated. Newman [39] explores ways to make model uncertainty more comprehensible, such as incorporating confidence intervals, conditional probabilistic predictions in natural language, and calibration techniques. These methods can be implemented organizationally through tools or systems designed for transparency, facilitating an intuitive understanding of the elements under development [40].

NIST also outlines guiding principles contributing to AI trustworthiness, serving as a regulatory and public policy framework. These principles include fairness, which necessitates the absence of harmful bias; accountability, identifying responsible parties for system errors or misconduct; and transparency, detailing the extent of information accessible to users for understanding AI systems' decision-making processes. Notably, NIST's principles are reflected in various AI regulatory proposals and design frameworks, such as those by the OECD, the EU, and local regulations like Japan's AI Governance [35].

Newman [39] emphasizes that without Transparency, even secure and reliable AI systems may fail to earn user trust. Accountability is closely linked to this, as identifying errors or misconduct in AI development, research, and implementation requires an understanding of how systems work and generate predictions. Transparency also involves disclosing the datasets used to train Narrow or Weak AIs (Artificial Intelligence), which often contain biases that must be recognized and mitigated by researchers; these

biases can influence the predictions made by AI platforms [38, 39].

Another concern is the transformative effects of technology on autonomy (Spence, 2021). For instance, in the Gig Economy, workers' independent decision-making is compromised by algorithms that influence their choices, coupled with surveillance issues [45], further encroaching on user autonomy. However, users and workers have also found ways to counter the effects on their autonomy by understanding how the systems work and resistance behaviors [37].

Considering the current discourse, the taxonomy of risks associated with AI models can be categorized into three main areas as suggested by NIST: technical and design attributes, perception of AI systems, and regulatory policy risks. Technical and design attributes include Accuracy, Reliability, Robustness (the system's defense against adversarial attacks), Security, and Resilience. Socio-technical attributes, which concern the perception of AI systems, encompass Explainability, Interpretability, Privacy, Safety, and the Absence of Bias. Regulatory policy risks involve Fairness, Accountability, and Transparency. According to Blackman [5], the most pressing risks that impact users and society are Privacy, Bias or discrimination, and Transparency; however, current AI models, including Generative AI, introduce additional risks with wide-ranging effects.

AI also presents risks related to social and environmental issues. For instance, Generative AI's role in spreading disinformation is exacerbated by the emergence of deep fakes [36] and biases within or produced by these models [4]. The extensive data required to train models, especially LLMs, Text-To-Image generators, and Video Generators like Sora, lead to significant energy and water consumption [31].

Risks beyond the established taxonomies can have profound political, social, and global consequences. Consider a scenario where actors use models like Sora to create and disseminate fake videos on social media, potentially damaging a political candidate's career, or where biased AI-generated evidence is used in legal proceedings against marginalized communities. The environmental impact of AI is already significant and is expected to increase as the energy demands for training large models like ChatGPT or Gemini escalate [61].

Moreover, the training of many AI models (particularly LLMs) is shrouded by human labor exploitation in the form of low-wage conditions. Therefore, it is important to understand that many elements of AI networks of associations always end with a human being [37], and in the case of AI models, these need people to supervise that they work correctly, even if these people are being paid low wages and suffering emotional harm by curating its content. As Alex Hanna [66] explained:

> "We know from reporting . . .that there is an army of workers who are doing annotation behind the scenes to even make this stuff work to any degree — workers who work with Amazon Mechanical Turk, people who work with [the training data company] Sama—in Venezuela, Kenya, the U.S., actually all over the world . . .They are doing the labeling, whereas Sam [Altman] and Emad [Mostaque] and all these other people who are going to say these things are magic—no. There are humans... These things need to appear as autonomous and it has this veneer, but there's so much human labor underneath it."

Exploring the taxonomy of risks associated with narrow AI can serve as both a methodological and reflective tool. However, the question arises: how can we integrate this with Stoic philosophy? Furthermore, how can these frameworks, when used together, provide a solid foundation for the debate concerning which risks hold greater significance—the immediate and avoidable risks posed by current AI technologies, or the speculative and, in any case, hardly preventable risks related to potential AGI and Super AI?

4 On fears of a super ai, public policy and the research and development of AI

Super AI and singularity raise concerns, and humans have a strong fear of the unknown. There is no more uncertain scenario than an AI singularity that may be capable of overcoming human capabilities, particularly intelligence. Carleton [9] defines fear of the unknown as an "…individual's propensity to experience fear caused by the perceived absence of information at any level of consciousness or point of processing." In the case of phenomena like AGI and Super AI, often described as singularity, this fear of the unknown is linked to the lack of sufficient information about what might happen if an artificial intelligence first becomes general and then turns into an entity whose capabilities human beings can hardly visualize. Thus, these fears create the worst-case scenario: a Singularity event that could potentially surpass human intelligence and control, leading to a future where AI's decision-making and abilities are beyond our understanding and prediction. This narrative of existential risk resonates with our deepest uncertainties about the trajectory of technological progress.

Unfortunately, this fear of the unknown sometimes takes away the importance of the most pressing problems regarding AI in its current narrow stage. For example, problems like deep fakes can have very plausible impacts on politics and elections, and while less believable deep fakes may lack credibility, they nonetheless exert a greater influence in undermining the legitimacy of the targeted

political figure [20]; in addition, deep fakes may be used to commit acts of extortion [67] or identity theft. In a way, the fears unknown surrounding AGI and Super AI may overshadow the immediate and tangible issues posed by Narrow AI.

Let us imagine a tentative scenario: what if Marcus Aurelius had succumbed to his fears regarding the unknown surrounding the German tribes and their battlefield environment? From a Stoic perspective, if Marcus Aurelius had yielded to his fears of the unknown concerning the Germanic tribes and their battlefield tactics, he would have abandoned the Stoic principle of focusing on what is within one's control. Stoicism would counsel him to accept the uncertainties of war while concentrating on his responses and strategies, which are within his power to command; in addition, Stoicism would advise him to prepare for and mitigate potential worst-case scenarios, and this approach can also be applied to the fears currently surrounding speculative scenarios regarding AGI and Super AI.

With a singularity event there are scenarios made from fear or doomsday representations of human extinction but also fears that tend to be more grounded. Among these plausible "if" scenarios in the framework of a singularity are social manipulation and the creation of new pathogens that can be harnessed against humanity [22]. This creates a sense of angst; still, all the plausible scenarios are still linked to if a Super AI emerges, and if it in turn decides to come to fruition those possible threats again linger into the unknown regarding the nature of that singularity scenario.

Films like The Terminator, Ex Machina, and Metropolis present a catastrophic scenario where an AI gains control of the world's nuclear weapons, leading to near-human extinction. Such stories, which often depict AI as a dominating force, have contributed to heightened public fear. These fears are influenced by Western religious traditions, as discussed by Jecker and Nakazawa [24], and may drive demand for stricter regulations to mitigate a potential singularity event, overshadowing the preventable and immediate concerns posed by current narrow AI technologies.

Additionally, there has been a sense of contradiction regarding some proponents of prevention and preparation regarding a singularity event. For example, in 2023 the CEO of OpenAI, Sam Altman talked at length in the media about the existential risk that a AGI and a Super AI posed [65]. Neverthless, Altam pushed to change certain details around the regulation of the so-called generative AI, to the European Union to soften its proposition on the regulation of these tools and applications [69].

How can Public Policy balance all these different factors regarding the pressing problems of narrow AI versus the fears and possibilities of an AGI or a singularity AI event? In this case, by balancing the things that are within or out of the control of regulations, as Spence [52] mentioned: "To avoid unhappiness, frustration, and disappointment, we, therefore, need to do two things: control those things that are within our control (our beliefs, judgments, desires, and attitudes) and be indifferent to those things which are not in our control (things external to us)." This is also related to Marcus Aurelius who claimed that there is no need for fear since it is in our power to inquire what ought to be done and those who follow reason are calm:

> What need have you of a hint or suggestion, when it is possible to see what ought to be done and, if you are conscious of that, kindly proceed on this path without turning back; but if you are not conscious of it, to suspend judgment and use the best men to advise you; or if some further points bar this advice, to go forward according to your present opportunities cautiously, holding fast to what seems to be just? For it is best to achieve justice, since, as you see, failure is to fail in this. The man who in everything follows the rule of Reason is at once master of his time and quick to act, at once cheerful in expression and composed (Marcus Aurelius, *Meditations,* X.12).

Are concerns about a technological Singularity well-founded in the context of today's technologies and algorithms? Consider two distinct perspectives on this issue. The first concerns the apprehension and challenges associated with a sentience that surpasses a robust sense of agency. From this viewpoint, as Véliz [55] posits, current AI algorithms are devoid of moral agency due to their lack of sentience: "Only beings capable of experiencing pain and pleasure can truly understand what it means to inflict pain or cause pleasure, and only those with this moral understanding can be considered moral agents."

The second, a more credible concern involves the feasibility of an AI that possesses understanding. Such an entity could emerge devoid of the moral constructs typically associated with humanity. In this scenario, it could proliferate unchecked, akin to a vine that overgrows and constricts. Ultimately, its absence of moral agency would be inconsequential, for its unchecked growth could still stifle us, stripping away our autonomy.

Even critics such as Véliz [55], who elucidates the absence of moral agency in algorithms, suggest that "there might come a time when AI becomes so sophisticated that robots might possess desires and values of their own." Nonetheless, while this debate is thought-provoking, it should not overshadow the pressing concerns presented by contemporary AI, including large language models (LLMs), given our limited capacity to mitigate such risks. Furthermore, technological advancement is not confined to any single nation or corporation: a prohibition in one region could simply be disregarded elsewhere. However,

should this stop efforts at prevention, which is the only thing that holds a candlelight of control under the clouds of uncertainty?

Echoing the sentiments of Musk and Hinton, the advent of AGI could be just a few years away, rendering any attempt to halt the myriad contributors to this development futile. In this context, the Stoic viewpoint is relevant: facing a seemingly inevitable occurrence, what purpose does worry serve? It is the immediate and controllable dangers, rather than the distant and inescapable ones, that ought to be prioritized in regulatory discussions and revisions to existing AI legislation, such as the AI Act in Europe [15–17]. Nonetheless, we cannot deny the danger that the current storm clouds may signal.

Perhaps the most interesting case to apply principles of Stoic Philosophy is the area of Research and development. This is because the design and research of AI systems of models is a crucial step in the prevention of harm or development of AI that works for the social good. However, it is important to highlight the distinction between two approaches to AI design, and also research and development; the first is AI for not bad, which focuses on the prevention of harms, risks, and elements that can have a negative use on these tools and platforms; the second is AI for good, which is centered more around the perspective of development of AI tools and systems that will help the flourishing of it users whether it is individuals or particular social group [5]; for example, by employing like Sousveillance Tools [45].

There is indeed a possibility of future developments in AI, but there is also the possibility of highly improbable events such as a quasar ray obliterating Earth. For now, our attention should remain firmly fixed on the development and implementation of this technology, the associated biases, and the human tendency to place more trust in technology than in other humans. Glikson and Williams-Woolley [19] aptly note, "Users are not always aware of the actual technological sophistication of AI; while in some cases, highly intelligent machines are operating at full capacity, in others, their capabilities may not be fully evident in their behavior."

Now, the speculation regarding the potential emergence of superintelligent AI was once a productive and inspiring topic within the AI community. For years, discussions surrounding how to ensure AI aligns with human rights, commonly referred to as the alignment problem, were largely motivated by speculations about the eventual emergence of superintelligence. Well-known texts like Stuart Russell's 2019 textbook emphasized the need for alignment to safeguard humanity against highly improbable events. Similar motivations are found in the works of Gabriel [18], Dewey [12], Risse [44], and Eckersley [13], as cited by LaCroix [29].

Besides the problems with alignment and AI risk, it is important to discuss another issue regarding Narrow AI that could also affect AGI. Many problems around so-called Generative AI, a sub-branch of Narrow AI, are related to data. For instance, certain image generator platforms have been found to recreate images they were originally trained on: "Stable Diffusion images with dataset similarity account for approximately 1.88% of our random generations" [50]. This raises ethical and legal questions about the originality and ownership of the generated images, as well as concerns about the misuse of data sources without appropriate attribution or consent. It also underscores that the content generated by these models, despite their complexity or apparent understanding, is still bound by the data on which they were trained. These problems could potentially spill into more complex Narrow AI models and even AGI systems.

Should the public, media, academia, public officials, and researchers ignore the risks of an AI singularity event? Let us draw an analogy of a meteorite impact: while there is a possibility that someday an asteroid may strike Earth, the current pressing problems—such as climate change—demand attention and resources. Where should these actors allocate their efforts and capital: to the preventable and immediate issues affecting human livelihood, or to a hypothetical and, in any case, hardly preventable catastrophic event? While the collision of a massive asteroid with Earth poses a genuine existential risk, the debate should not overshadow other urgent matters.

Based on Stoic philosophy, the answer may not be as binary as a simple yes or no. The public, policymakers, and AI engineers must balance the current and immediate risks of narrow AI with the problem of the "lazy argument" discussed by the Stoics: the tendency to not act because something is unlikely to happen or does not pose an immediate risk. This balanced approach ensures that while preparing for distant and less probable catastrophic events, we do not neglect the pressing issues that currently affect human livelihood. Like the preventive measures in place to detect and monitor asteroids and meteorites that might threaten our planet, we must address the immediate risks posed by AI while remaining vigilant about potential future threats.

5 Discussion on the complexities of AI fears and opportunities

The term "Singularity" is often used as a sophisticated way of expressing uncertainty about the future, particularly regarding the advent of Artificial General Intelligence (AGI) or Super AI. This concept frequently incites fear due to its association with the unknown—a sentiment amplified by Western narratives that often depict AGI or Super AI as harbingers of doom. Such fears are fueled by public figures like Elon Musk, who have publicly expressed their concerns, and by a deeply ingrained human apprehension of the unknown.

The notion of the Technological Singularity, as conceptualized by futurists like Ray Kurzweil [27] and popularized by individuals including Elon Musk [64], describes a hypothetical point in the future where AI and other technologies catalyze a swift and profound acceleration of human progress. This could potentially lead to the creation of a superintelligent entity surpassing human intellect. In this context, the Singularity is referred to as a "Singularity event"—a moment filled with uncertainties that may or may not occur. As described by Reji, Sangeetha, and Silpa [23], technological singularity is a theoretical situation linked to the rise of artificial general intelligence, also known as "strong AI."

The debate over whether the emergence of a Singularity Event is within our control encompasses various perspectives. On one hand, the development of advanced AI systems is a human endeavor, subject to the decisions of governments, organizations, and researchers worldwide, who engage in AI research and development under diverse regulatory frameworks. On the other hand, factors beyond any single entity's control could shape AI's trajectory. For instance:

1. International collaboration and competition in technology mean that regional AI regulations may be bypassed by progress in other areas.
2. The pace of technological advancement is determined by a myriad of elements, including scientific breakthroughs, economic motives, and societal shifts, which may elude complete control by any governing body.
3. The unintended consequences of developing advanced AI systems could unexpectedly hasten or delay a Singularity event, and these outcomes are often unpredictable.

In addition, it is important to work within four frameworks regarding AGI: immediacy, preventability, likelihood, and uncertainty. *Immediacy* refers to the present and pressing nature of certain events or situations. Stoic philosophers like Seneca and Marcus Aurelius emphasized the importance of focusing on the present moment and not being overly concerned with the distant future or past, believing that wisdom involves making the best use of the present. "[…] practise only to live the life you are living, that is the present, then you will have it in your power at least to live out the time that is left until you die, untroubled and with kindness and reconciled with your own good Spirit" (Marcus Aurelius, *Meditations*, XII.3). From this belief system, their views support the priority we assign to the risks of Narrow AI over those of AGI.

Preventability is another critical aspect, broken down into understandable fragments regarding the dangers of AI, by dealing with whether an event or situation can be avoided or mitigated. Seneca acknowledged that while some things are within our control, many are not, with his teachings often revolving around focusing on what one can control and accepting what one cannot (*On Providence*) [49]. As Hinton claims, while we know that climate change is primarily caused by fossil fuels, we don't know how the singularity may emerge, stating: "I wish it was like climate change, where you can say, 'Stop burning carbon.' There isn't a simple recipe like that for AI" [70].

Likelihood pertains to the chances of events occurring. Seneca recognized that life is filled with uncertainties and unpredictable events, and he advised that it is important to prepare oneself mentally for all possibilities, thereby reducing the impact of unforeseen events. "Our minds should be sent forward in advance to meet all problems, and we should consider, not what is wont to happen, but what can happen. For what is there in existence that Fortune, when she has so willed, does not drag down from the very height of its prosperity?" (Seneca, *Letter to Lucilius*, XCI.4) [47]. Now, given constant innovation, the probability of an AI singularity event emerging in the next decades is not completely out of the realm of possibility.

Finally, *uncertainty* refers to the lack of predictability regarding future events. Seneca's philosophy teaches us to embrace uncertainty and remain calm and composed in the face of it. The philosopher argued that our mental attitude towards uncertainty can significantly affect our well-being. In the case of AGI, the advice from Stoic philosophy is clear: practice the virtues of tranquility and temperance. As Seneca stated, "It was a great deed to conquer Carthage, but a greater deed to conquer death" (*Letter to Lucilius*, XXIV.10) [46]. We find his description of tranquility and how to cure anxiety and worry in the dialogue *De Tranquillitate Animi* (*On Tranquility of Mind*, 2.4) [49].

In the realm of global AI development competition, efforts to regulate or prevent a Singularity by one nation or coalition could be neutralized by advancements elsewhere, such as in China. The potential rise of Super AI would be influenced by a complex mix of technological, societal, and economic factors, making it difficult for any one party to fully govern this evolution. Moreover, "it's unlikely that the AI community, governments, and corporations that control the large research budgets, especially the multi-billion-dollar budgets of Big Tech companies, will respond to *the gorilla problem* by stopping AI research" [52].

Within this context, it is feasible to explore mediation strategies for discussions on urgent AI regulations. The authors have created a conceptual map (Fig. 1) that visually depicts the intricate relationships between AI's various stages and societal impacts, underscoring the importance of addressing ethical aspects like fairness, privacy, and safety, in addition to technical risks such as accuracy and robustness. The map's incorporation of Stoicism suggests a philosophical stance on AI ethics, promoting a balanced approach to challenges and offering the epistemological

tools to navigate the complex AI landscape, from Narrow AI's practical concerns to the existential questions posed by Super AI and Singularity.

The conceptual map organizes a network of associations that begins with the foundational elements: AI, Narrow AI, AGI, Super AI, and Singularity. It then categorizes the challenges or risks into another set of associations: Unknowns, Fears, Plausible Fears; Knowns, Risks, and the subcategories of risk. Lastly, it incorporates the philosophy of Stoicism to comprehend and contextualize the significance of these risks. This is done to anchor the discussion on the comparative risks of Narrow AI and the potential Singularity event, as well as to emphasize the necessity for urgent and preventative public policy.

Based on the conceptual map and our earlier discussion on the problem of Narrow AI, it is important to also highlight what other authors and institutions mention about the risks associated with AGI and Super AGI. Yudkowsky [76] argues that we are not prepared for the challenges of AGI, saying: "There is no plan. Progress in AI capabilities is running vastly, vastly ahead of progress in AI alignment or even progress in understanding what the hell is going on inside those systems." Additionally, Yudkowsky explains that the problem with AGI research, or even self-aware AI, is that researchers often do not fully understand what they are doing. He emphasizes, "This is alarming not just because of the moral implications of the 'self-aware' part, but because being unsure means you have no idea what you are doing and that is dangerous, and you should stop" (ibid).

What are the tentative risks regarding AGI? As mentioned in this essay, perhaps the biggest risk is human extinction or a catastrophic event where we forfeit our agency to these superintelligent systems. Moreover, there are bellicose applications, also known as "necroware." Andrade [2] explains, "Necroware is an artificial superintelligence whose subjectivity and improvement process is aimed at perfecting war and information skills to initiate a new techno-biological era through violence."

Considering the current landscape, the focus on AI regulation and design is shifting towards addressing the immediate challenges posed by Narrow AI. This prioritization helps to clarify the debate over which AI-related issues are most pressing. Researchers, ethicists, and regulators are cognizant of AI's problems, including emerging phenomena like deep fakes. However, the concept of a Singularity event remains shrouded in uncertainty, making it difficult to predict or prepare for; therefore, it is pragmatic to concentrate on Narrow AI issues while simultaneously, and with less urgency,

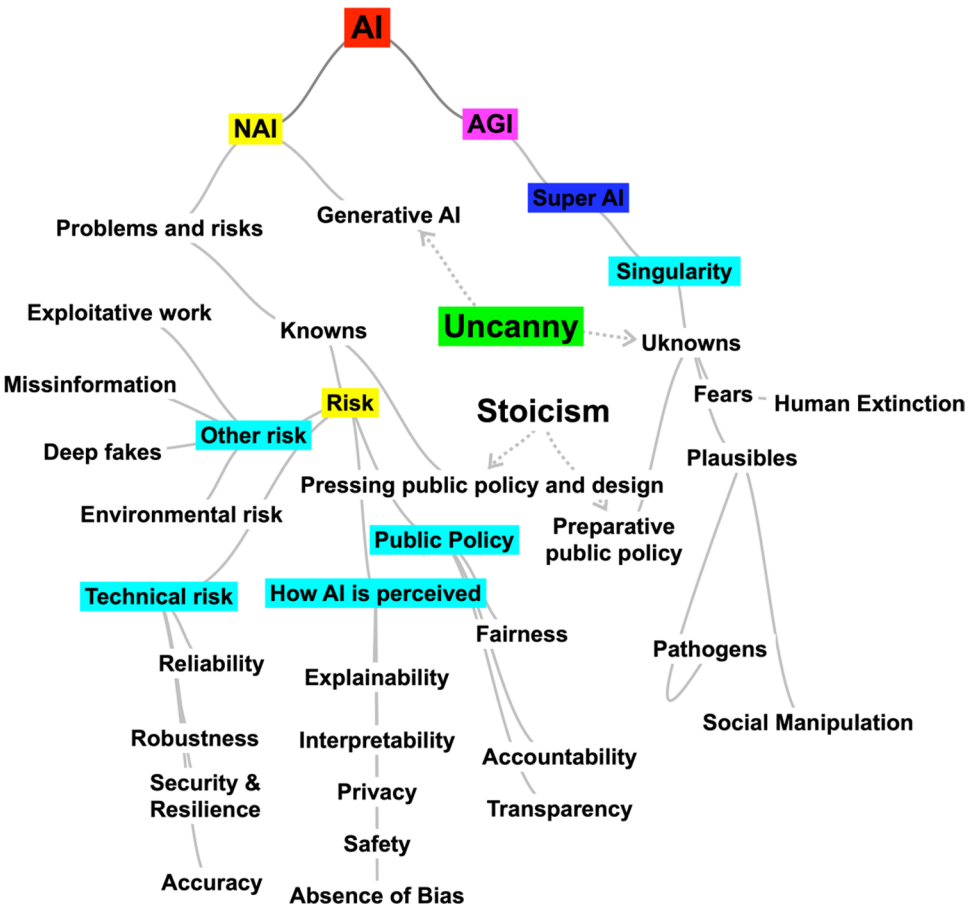

Fig. 1 The conceptual map that describes the associations of risk regarding Narrow AI and Super AI

developing strategies to mitigate the potential risks of Super AI, and Seneca's perspective on uncertainty is relevant here:

> "I will tell you: that perfect man, who has attained virtue, never cursed his luck, and never received the results of chance with dejection; he believed that he was citizen and soldier of the universe, accepting his tasks as if they were his orders. Whatever happened, he did not spurn it, as if it were evil and borne in upon him by hazard; he accepted it as if it were assigned to be his duty. "Whatever this may be," he says, "it is my lot; it is rough and it is hard, but I must work diligently at the task." (Seneca, *Letters to Lucilius*, CXX.12) [48].

A potential framework for addressing both current Narrow AI and hypothetical Super AI challenges revolves around the concept of designing AI for no harm and AI for good. Toxtli et al. [53] suggest that artistic creativity can lead to innovative ideas that push the boundaries of what is currently possible, offering new perspectives on AI design. This creative approach could also provide a pathway to explore Narrow AI systems for positive outcomes and begin to consider the risks associated with a Singularity event.

Furthermore, incorporating diverse perspectives into AI design can greatly enrich the process, emphasizing the importance of public and user participation. As Toxtli et al. [53] advocate, including gig workers in "creative artistic co-design sessions" and integrating insights from ethical philosophy and sociology can offer a more comprehensive approach to designing Narrow AI and preparing for potential Singularity scenarios. This inclusive strategy may also help address concerns about which AI category—Narrow AI or AGI—requires immediate regulation.

It is essential to recognize and address the influence that current AI has on our control and its impact on daily life. While the focus may currently be on the risks associated with AI and AGI, Spence [52] notes that even if we resolve the technical control issues of AGI, the broader problem of meta-control will remain. This challenge will persist as long as large tech corporations dominate AI technologies. But does this mean that Stoics would advise against taking any action?

Yudkowsky [58] suggested that humans tend to anthropomorphize, which implies caution is needed when developing general AI. When addressing crucial and complex problems, it is important not to assume that a general intelligent and moral agent will necessarily align with human notions of intelligence. For example, intelligence in nature can differ significantly from human intelligence, as illustrated by the behavior of an octopus (Guerra, n.d.). Humans fear a singularity event partly because we project human characteristics onto such entities, assuming they would replicate the harmful behaviors humans inflict on other beings. However, it is also a mistake to believe that an AI will inherently be friendly, as this assumption could lead to a path toward global catastrophe [58].

Without a doubt, this is not an excuse for the lack of preventive action, as the Stoics understood when they responded to the Lazy Argument, also known as the Idle Argument. The Lazy Argument was probably proposed by the ancient philosopher Carneades, a prominent Academic Skeptic. Carneades (214/213–129/128 BCE), known for his critical stance towards Stoicism and other dogmatic philosophies of his time, may have used the Lazy Argument to challenge the Stoic belief in determinism and fate. And the most popular form of this argument, presented by Cicero (*De Fato*, XII, 28–29) [10] argues that if destiny dictates your recovery from an illness, you will recover regardless of whether you seek a doctor's help. Conversely, if destiny dictates that you will not recover, a doctor's intervention will not change the outcome. Since one outcome is already determined by fate, seeking medical assistance is futile.

If everything is fated to happen, then what will happen will occur regardless of what we do. Therefore, our actions would not matter because the outcome would be predetermined, making it pointless to make any effort or take any action since fate will determine the outcome regardless. Nonethless, the Stoics distinguished between different types of causes: the overarching, deterministic forces that govern the universe (fate) and the immediate, proximate causes that contribute to outcomes (human actions). A river's course flowing to the sea (representing the fated outcome) includes various tributaries (representing human actions). Even if the river is destined to reach the sea (the emergence of the singularity), the exact path it takes is shaped by these tributaries.

Removing a tributary (a human action aimed at aligning AI with human values) would alter the river's path, though it would ultimately still reach the sea. This analogy highlights that not all scenarios regarding the potential emergence of singularity are equal. Consequently, even if research into this technology is somewhat speculative, it does not necessarily lead to an AGI that could doom humanity. Research efforts can be shaped and influenced by precautionary measures, including regulation, self-regulation, and designing and studying both narrow and general AI systems with a focus on ethical considerations.

This is like individuals who eat healthily and exercise to avoid the prospect of a short life while also addressing their immediate concerns in the present. A balanced approach is crucial in both current (Narrow) and potential (General) AI scenarios. It may seem reasonable to sacrifice sleep and proper nutrition in youth due to pressing matters like romance or finishing an academic essay to advance a career. Still, it is also rational to maintain one's health and well-being in the long term to avoid significant risks: striking a balance is both possible and necessary.

Certainly, it is essential to ground the discussion by comparing risk perspectives for both narrow and general AI. Focusing solely on a speculative singularity scenario might lead to the neglect of current issues and regulations related to narrow AI in public and policy discussions, including aspects of regulation and self-regulation. Conversely, concentrating only on current narrow AI problems could overshadow the potential risks and consequences associated with the emergence of General AI. Therefore, a comprehensive approach is needed, addressing both the immediate challenges of narrow AI and the long-term implications of General AI, to ensure a balanced and forward-thinking regulatory framework.

A simple logical form of the argument [7] is a complex constructive dilemma familiar to the Stoics:

a. If A, then B.
b. If C, then D.
c. Either A or C.
d. Therefore, either B or D.

The conclusion presents a disjunction and does not assert idleness. As [34] (p. 371) notes, "Hence, we can conclude that either the argument is not complete or that the suggested inference form is not proper, since, at this stage, it does not appear to be a validly inferred conclusion." While we can align with Cicero's formulation and reflect idleness in the conclusion, this would render the first premise clearly false:

e. If the Singularity will emerge, then it makes no difference whether we try to align it to human values or not.
f. Similarly, if the Singularity will never appear, then it makes no difference whether we try to align it to human values or not.
g. Either the Singularity will emerge or not.
h. Therefore, there is no point in trying to align it to human values.

How can we be so sure that our efforts to align with AGI will be pointless? After all, both narrow AI and AGI are human-made and can be influenced by our actions. The real reason behind the rejection of moratoriums and regulations on AGI appears to be the associated costs, including the expenses of surveillance and the potential losses from hindering technological innovation.

The problem is the distribution of scarce resources between the regulation of narrow AI and General AI. Because the cost of avoiding the immediate risks of narrow AI for democracy, employment and crime are already very high. And it would also be costly to try to forcibly align with human values all narrow AI projects that could potentially jump to the General AI. Thus, a weighting or trade-off is necessary. These may consist of allocating part of the budget of government agencies and ministries to defend democracy, employment, and public safety in face of narrow AI's risks and, more modestly, as Hinton [70] proposes, forcing AI developers to allocate the same amount of money they invest in new AI innovations in alignment research, alignment of AI with human values.

While the former involves government expenditures, the latter pertains to costs associated with regulatory norms. Specifically, this distinction separates state-funded investments from corporate expenditures aimed at meeting official requirements. Consequently, this distinction parallels the difference between funding for applied research and basic research. In this context, the singularity is not an object of applied research in the same way that narrow AI represents a tangible risk requiring immediate attention and management; therefore, just as funding priorities differ between applied and basic research, so should our approach to addressing the risks of singularity versus those associated with narrow AI.

To address both the schism problem and the tendency to focus disproportionately on the risks of AGI—often exacerbating fear—while neglecting current issues related to Narrow AI, we must consider notions of action and prevention. Imagine living in a country that faces significant crime in its major cities and experiences occasional massive earthquakes. In this scenario, the sensible approach would be to act on what you can control, such as implementing policies and improving policing to address crime. Conversely, the "lazy argument" would suggest ignoring the problem of earthquakes because they are infrequent and unpredictable, with even the best scientists unable to forecast their occurrence; thus, while it is crucial to prepare for and mitigate immediate and controllable risks, we should not disregard potential long-term threats simply because they are less predictable.

Here, the focus would be on addressing the immediate problem of crime, which is currently affecting society. While it might seem less urgent to worry about an earthquake that cannot be accurately predicted, the question arises: what if media, academia, and policy discussions neglected to address crime because they were preoccupied with the potential catastrophic consequences of an earthquake? How do we find the right balance?

The control problem offers guidance for both action and prevention. In an earthquake-prone country, governments would address the immediate issue of crime while also preparing for potential severe earthquakes. This preparation includes not only implementing strict building codes and conducting earthquake drills but also establishing educational programs to inform the public and children about what to expect and how to respond, and fostering collaboration among the private sector, government, and academia. Similarly, while we may not yet know how to build an AI with

human-level intelligence and thus cannot be certain when it might emerge, as Yudkowsky et al. [59] noted, "we can't rule out unforeseen advances." Thus, just as with earthquake preparedness, we must strive to mitigate the risks of a potential Singularity without neglecting the pressing and tangible human rights issues associated with current Narrow AI.

Therefore, it is prudent to implement policies that address the current problems of AI while also preparing preventive measures for a potential Singularity event. Although not all countries may act on AGI prevention and some might disregard international regulations, a balanced approach that combines immediate action on current AI issues with proactive measures for future risks provides a more comprehensive and responsible strategy. For instance, human gene editing, despite being a highly controversial ethical issue, offers a valuable comparison because the debate surrounding gene editing highlights the necessity for a balanced approach: while it is crucial to address immediate ethical and safety concerns, it is equally important to develop proactive guidelines and regulations for future advancements. Similarly, in AI, we must address current challenges while also preparing for potential future scenarios to ensure responsible and effective management.

6 Conclusions

We have argued that the distinction between the treatment of narrow AI and AGI is akin to the distinction between funding applied research and funding basic research. While some resources should be allocated to speculative and exploratory topics (basic research), most research efforts should be directed towards solving immediate human problems (applied research). This distinction underpins our practical proposal: the risks associated with narrow AI should be addressed with a standard government budget, similar to how issues such as transparency, democracy, and crime prevention are funded. Conversely, the risks of AGI may be mitigated through investments resulting from regulatory measures imposed on companies. This approach aligns with Hinton's proposal that companies should be required to allocate a percentage of their investment in innovation towards ensuring alignment with human values.

Implementing policies to address current AI problems while also creating preventive measures for a potential singularity event is prudent. Not all countries may act on AGI prevention, and some actors may choose to ignore international regulations. However, a balanced approach that involves immediate action on current AI issues and proactive measures for future risks can provide a more comprehensive and responsible strategy. This approach can help prevent potential issues and establish guidelines for dealing with actors who ignore AGI regulations, like the Treaty on the Non-Proliferation of nuclear weapons. This is especially important given that "the estimated arrival time of AGI… is close by in the next quarter century to mid twenty-first century…before any planetary global warming emergency takes full effect" [33].

It is also important to draw from the taxonomies of risk for both Narrow AI and AGI. Beyond self-regulation, we need robust theoretical and legal frameworks to enforce ethical research of these systems. By focusing on what we can control now, we can better understand and mitigate the risks posed by corporations, which "through the various activities of their platforms, exert enormous monopolistic global power over all sectors of society, social, political, and financial" [52]. While it is complicated and arguably impossible to halt AI research altogether, international frameworks have successfully addressed and mitigated other complex issues. For example, many countries have banned human cloning research [57], while still allowing for the benefits of gene therapy. This demonstrates that it is possible to manage the risks of emerging technologies while still harnessing their benefits.

Another approach to preventing a catastrophic scenario and exercising control over AI research, implementation, and design—while also addressing the current problems of Narrow AI—is to embed ethical frameworks in AI development. As Spence [52] suggests, we should "act wisely and design the right values in AGI agents, including love, to be not only intelligent but more importantly to be wise." By incorporating both international and self-regulation, and understanding local sensibilities and worldviews on ethics, we can foster a more responsible and beneficial AI landscape.

We must underline that the advent of AI brings both remarkable opportunities and formidable challenges, sparking concerns about its environmental footprint, job displacement, privacy violations, biases, and the spread of disinformation. Amidst these issues, Stoicism provides a pertinent framework for navigating the complexities of the AI era. This ancient philosophy, centered on self-control, resilience, and acceptance, teaches focusing on what we can control and accepting what we cannot [52], as it suggests that we address AI's preventable and immediate challenges, like its environmental impact, through moderation and self-discipline, by optimizing algorithms or investing in green energy.

Spence [52] elaborates on Stoic ethics, stating that we should prioritize developing virtues such as courage, moderation, justice, and practical wisdom, which are essential for human flourishing. When aspects of technology like accuracy, transparency, privacy, and control are beyond our reach, Stoicism offers two paths: either to treat these as indifferent since they don't affect our well-being or to strive to bring them under societal control due to their potential impact on our lives.

Stoicism thus advises us to tackle AI's immediate issues while acknowledging the limits of our influence, especially regarding speculative events like Singularity. The rise of superintelligent AI, shaped by various factors, may be out of our hands. Instead of yielding to fear, Stoicism encourages us to concentrate on our present thoughts, actions, and choices. This perspective doesn't mean we ignore the risks of a Singularity event, but rather that we focus on controllable aspects of current Narrow AI, which profoundly affects society; thus, the discussion should revolve around regaining control over AI through regulation, design, and research; in addition to measures to prevent exploitation and environmental harm from AI development and deployment.

Another proposal akin to the Stoic perspective is to design AI systems from the standpoint of "minimum virtuous products," a term mentioned by Taneja [73]. This approach to design, research, and development can be applied to both Narrow AI and AGI research. It is crucial to understand that "if innovation is to survive into the twenty-first century, we need to change how companies are built by changing the questions we ask of them." One of the dangers in the current IT sector, including AI, is the quasi-ethos of "Move fast and break things," which has already resulted in numerous human rights problems with current public AI models; moreover, this approach can lead to unexpected and unwanted consequences, such as the hallucinations seen in many transformer models like ChatGPT.

It is worth noting that the Stoic perspective has recently become both overused and over-simplified in various societal contexts. Critics have pointed this out [25]. Yet, this enduring tradition remains one of the most crucial tools in the Western philosophical arsenal for engaging in the debate surrounding AI risks and their potential policy implications.

AI is a complex set of technologies, and its current and potential risks have led to movements like the AI Moratorium [60], which called for a pause on AI research. However, a profound divide has emerged among different sectors: some advocate for urgent regulation and preventive measures for Narrow AI, while others focus on the potential catastrophic scenarios posed by a plausible Super AI. This deep chasm creates noise in the discussions about these problems. By analyzing the ongoing academic and public debates between these two camps—those emphasizing the immediate issues of Narrow AI and those concerned with the long-term risks of General AI—we can turn to Stoicism and the control problem for guidance.

This essay applies Stoic philosophy and the principle of control to the discourse on Narrow AI versus the theoretical Singularity AI event. While it does not delve into empirical applications, such as how Stoicism could inform the creation of improved AI systems and models, it underscores the importance of integrating these philosophical concepts into future research and academic inquiry. Moreover, by applying some perspectives of Stoic philosophy to an empirical paradigm, this work highlights the significance of this branch of academia as an ethical and analytical aid in other areas of research.

Finally, Stoicism provides a relevant and actionable framework for understanding the AI era. By adhering to the Stoic principle of focusing on what we can control, we can more effectively evaluate the most pressing risks associated with both current and future AI models. Moreover, grounding our discussions in Stoic philosophy helps address challenges in design, education, and public communication, and guides the formulation of public policy. This approach emphasizes the importance of concentrating on risks within our control, rather than becoming overwhelmed by those beyond our influence.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43681-024-00548-w.

Author contributions Not applicable.

Funding Open access funding provided by Universidad Autonoma Metropolitana (BIDIUAM).

Data availability Not applicable.

Declarations

Conflict of interest No possible point of interest influenced the research and the writing of this work.

Informed consent Not applicable.

References

1. AI Act.: Shaping Europe's digital future (2024). https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Accessed July 2024
2. Andrade, R.: Problemas filosóficos de la inteligencia artificial general: ontología, conflictos ético-políticos y astrobiología. Argumentos de razón técnica **26**, 275–302 (2023). https://doi.org/10.12795/Argumentos/2023.i26.10
3. Aljaber, S., et al.: International journal of engineering research and applications.** *International journal of engineering research and applications*, *12*(12), 52–57 (2022). Retrieved

from https://www.ijera.com/papers/vol12no12/G12125257.pdf. Accessed July 2024
4. Aničin, L., Stojmenović, M.: Bias analysis in stable diffusion and midjourney models. In: Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, pp. 378–388. Springer Nature, Switzerland (2023). https://doi.org/10.1007/978-3-031-35081-8_32
5. Blackman, R.: Ethical machines: Your concise guide to totally unbiased, transparent, and respectful AI. Harvard Business Press, Boston (2022)
6. Bartneck, C., Lütge, C., Wagner, A., Welsh, S.: What is ai? In: Bartneck, C., Lütge, C., Wagner, A., Welsh, S. (eds.) An introduction to ethics in robotics and AI, pp. 5–16. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-51110-4_2
7. Bobzien, S.: Determinism and freedom in Stoic philosophy. Oxford University Press, Oxford (1998)
8. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: early experiments with GPT-4. ArXiv (Cornell University) (2023). https://doi.org/10.48550/arxiv.2303.12712
9. Carleton, R.N.: Fear of the unknown: one fear to rule them all? J. Anxiety Disord. **41**, 5–21 (2016). https://doi.org/10.1016/j.janxdis.2016.03.011
10. Cicero, M.T.: De Fato, Latin. Aris and Phillips Classical Texts (1991)
11. Cicero, M.T.: On divination. Oxford University Press, Oxford (2006)
12. Dewey, D.: Learning what to value. In artificial general intelligence: 4th international conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings 4, pp. 309–314. Springer Berlin Heidelberg, (2011)
13. Eckersley, P.: Impossibility and uncertainty theorems in AI value alignment (2019)
14. Epictetus: Discourses, Fragments, Handbook. Oxford University Press, Oxford (2014). (**Hard R. (translator)**)
15. European Parliament: Artificial Intelligence act briefing. European parliamentary research service (2021). https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI%282021%29698792_EN.pdf. Accessed July 2024
16. European Parliamentary Research Service: General-purpose artificial intelligence (2023). Retrieved from https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/745708/EPRS_ATA(2023)745708_EN.pdf. Accessed July 2024
17. EU AI Act: first regulation on artificial intelligence | Topics | European parliament. Topics | European Parliament (2023). https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed July 2024
18. Gabriel, I.: Artificial intelligence, values, and alignment. Mind. Mach. **30**(3), 411–437 (2020)
19. Glikson, E., Woolley, A.W.: Human trust in artificial intelligence: review of empirical research. Acad. Manag. Ann. **14**(2), 627–660 (2020). https://doi.org/10.5465/annals.2018.0057
20. Hameleers, M., Van Der Meer, T.G., Dobber, T.: Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deep fakes. Comput. Hum. Behav. **152**, 108096 (2024). https://doi.org/10.1016/j.chb.2023.108096
21. Helmore, E.: "We are a little bit scared": OpenAI CEO warns of risks of artificial intelligence. The Guardian (2023). https://www.theguardian.com/technology/2023/mar/17/openai-sam-altman-artificial-intelligence-warning-gpt4
22. Hendrycks, D., Mazeika, M., Woodside, T.: An overview of catastrophic AI risks (arXiv:2306.12001). arXiv. http://arxiv.org/abs/2306.12001 (2023)
23. Issac, R.M., Sangeetha, K.S., Silpa, A.S.: Technological Singularity in Artificial Intelligence. Unpublished (2020). https://doi.org/10.13140/RG.2.2.32607.84646
24. Jecker, N.S., Nakazawa, E.: Bridging east-west differences in ethics guidance for AI and robotics. AI **3**(3), 764–777 (2022). https://doi.org/10.3390/ai3030045
25. Karl, J.A., Verhaeghen, P., Aikman, S.N., Solem, S., Lassen, E.R., Fischer, R.: Misunderstood stoicism: the negative association between Stoic ideology and well-being. J. Happiness Stud. **23**(7), 3531–3547 (2022)
26. King, M.R., chatGPT.: A conversation on artificial intelligence, chatbots, and plagiarism in higher education. Cell. Mol. Bioeng. **16**(1), 1–2 (2023). https://doi.org/10.1007/s12195-022-00754-8
27. Kurzweil, R.: La Singularidad está cerca: Cuando los humanos transcendamos la biología. Lola Books (2015)
28. Kuusi, O., Heinonen, S.: Scenarios from artificial narrow intelligence to artificial general intelligence—reviewing the results of the international work/technology 2050 study. World Futur. Rev. **14**(1), 194675672211016 (2022). https://doi.org/10.1177/19467567221101637
29. LaCroix, T.: Artificial intelligence and the value alignment problem, Toronto: Canadian philosophical association meeting, (2023)
30. Liu, N., Brown, A.: AI increases the pressure to overhaul the scientific peer review process. Comment on "Artificial Intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened." J. Med. Internet Res. **25**, e50591 (2023). https://doi.org/10.2196/50591
31. Luccioni, A.S., Hernandez-Garcia, A.: Counting carbon: a survey of factors influencing the emissions of machine learning. arXiv. http://arxiv.org/abs/2302.08476 (2023)
32. Lund, B.D., Wang, T.: Chatting about ChatGPT: How may AI and GPT impact academia and libraries? Libr. Hi Tech News **40**(3), 26–29 (2023). https://doi.org/10.1108/LHTN-01-2023-0009
33. Macey-Dare, R.: How soon is now? predicting the expected arrival date of AGI-Artificial General Intelligence. Available at SSRN: https://ssrn.com/abstract=4496418 (2023)
34. Marko, V.: Looking for the lazy argument candidates (1). Organon F **18**(3), 363–383 (2011)
35. Ministry of Economy, Trade and Industry (METI): AI Governance in Japan Ver. 1.1 Report from the Expert Group on How AI Principles Should be Implemented (2021). Retrieved from https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_8.pdf
36. Moreno, F.R.: Generative AI and deepfakes: a human rights approach to tackling harmful content. Int. Rev. Law Comput. Technol. (2024). https://doi.org/10.1080/13600869.2024.2324540
37. Morton, J.L.: On actor-network theory and algorithms: chatGPT and the new power relationships in the age of AI. AI Ethics (2023). https://doi.org/10.1007/s43681-023-00314-4
38. Morton, J.L.: On inscription and bias: data, actor network theory, and the social problems of text-to-image AI models. AI Ethics (2024). https://doi.org/10.1007/s43681-024-00431-8
39. Newman, J.: A taxonomy of trustworthiness for artificial intelligence: connecting properties of trustworthiness with risk management and the AI lifecycle. Center for long-term cybersecurity (2023)
40. NIST.: Draft -Taxonomy of AI Risk. Recuperado el 29 de noviembre de 2023, de 4 (2021)
41. O'Keefe, T.: Ancient theories of freedom and determinism. In: Zalta, E.N. (ed.) The Stanford encyclopedia of philosophy (spring 2021 edition). https://plato.stanford.edu/archives/spr2021/entries/freedom-ancient/ (2021)
42. O'Neil, C.: Weapons of math destruction: how big data increases inequality and threatens democracy (1. an ed., Vol. 1). Crown Publishing Group, NewYork (2016)

43. Pigliucci, M.: What is and is not in our power: a response to christian Coseru. Reason Pap. **40**(2), 19–33 (2018)
44. Risse, M.: Human rights and artificial intelligence: an urgently needed agenda. Hum. Rights Q. **41**, 1–16 (2019)
45. Santos, M.D.L., Do, K., Muller, M., Savage, S.: Designing sousveillance tools for gig workers. arXiv. https://doi.org/10.48550/ARXIV.2403.09986 (2024)
46. Seneca, L.A.: Moral letters to Lucilius (Epistulae morales ad Lucilium), Mott Gummere, R. (translator), Loeb Classical Library, vol. 1 (1917)
47. Seneca, L.A.: Moral letters to Lucilius (Epistulae morales ad Lucilium), Mott Gummere, R. (translator), Loeb Classical Library, vol. 2 (1920)
48. Seneca, L. A.: Moral letters to Lucilius (Epistulae morales ad Lucilium), Mott Gummere, R. (translator), Loeb Classical Library, vol. 3 (1925)
49. Seneca, L. A.: Hardship and Happiness, Fantham et al., (translators). University of Chicago Press (2014)
50. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion Art or Digital Forgery? Investigating data replication in diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 6048–6058. https://openaccess.thecvf.com/content/CVPR2023/papers/Somepalli_Diffusion_Art_or_Digital_Forgery_Investigating_Data_Replication_in_Diffusion_CVPR_2023_paper.pdf(2023).
51. Sora: Creating video from text. (n.d.-c). https://openai.com/sora. Accessed July 2024
52. Spence, E.: Stoic philosophy and the control problem of AI technology: Caught in the web. Rowman & Littlefield (2021)
53. Toxtli, C., Suri, S., Savage, S. Quantifying the invisible labor in crowd work. arXiv. https://doi.org/10.48550/ARXIV.2110.00169 (2021)
54. Van Dijck, J., Poell, T., De Waal, M.: The platform society. In: Public values in a connective world. Oxford University Press, Oxford (2018)
55. Véliz, C.: Moral zombies: Why algorithms are not moral agents. AI Soc. **36**(2), 487–497 (2021). https://doi.org/10.1007/s00146-021-01189-x
56. Vold, K., Harris, D.R.: How does artificial intelligence pose an existential risk? In: Veliz, C. (ed.) The Oxford handbook of digital ethics. Oxford University Press, Oxford (2021). https://doi.org/10.1093/oxfordhb/9780198857815.013.36
57. Wheat, K., Matthews, K.: World human cloning policies. Rice university's baker institute. Retrieved from https://www.ruf.rice.edu/~neal/temp/ST%20Policy/index/SCBooklet/World.pdf (2014)
58. Yudkowsky, E.: Artificial Intelligence as a positive and negative factor in global risk. In: Yudkowsky, E. (ed.) Global catastrophic risks. Oxford University Press, Oxford (2008). https://doi.org/10.1093/oso/9780198570509.003.0021
59. Yudkowsky, E., Salamon, Shulman, C., Kaas, S., McCabe, T., Nelson, R.: Reducing long-term catastrophic risks from artificial intelligence. In MIRI MACHINE INTELLIGENCE RESEARCH INSTITUTE. The singularity institute, San Francisco, CA (2010)

Hemerographic references

60. AI moratorium (2023). https://moratorium.ai/. Accessed July 2024
61. Burke, J.: Assessing the environmental impact of large language models. Enterprise AI (2023). https://www.techtarget.com/searchenterpriseai/tip/Assessing-the-environmental-impact-of-large-language-models. Accessed July 2024
62. Field, H.: OpenAI's Sam Altman reverses threat to cease European operations. CNBC (2023). https://www.cnbc.com/2023/05/26/openai-ceo-sam-altman-reverses-threat-to-cease-european-operations.html. Accessed July 2024
63. Guerra, C. (s. f.). Why the octopus brain is so extraordinary | Smithsonian Ocean. From July 2024, source https://ocean.si.edu/ocean-life/invertebrates/why-octopus-brain-so-extraordinary. Accessed July 2024
64. Isaacson, W.: Inside Elon Musk's struggle for the future of AI. TIME (2023). https://time.com/6310076/elon-musk-ai-walter-isaacson-biography/. Accessed July 2024
65. Kelly, S. Sam Altman warns AI could kill us all. But he still wants the world to use it. CNN (2023). https://edition.cnn.com/2023/10/31/tech/sam-altman-ai-risk-taker/index. Accessed July 2024
66. Loizos, C.: TechCrunch is part of the Yahoo family of brands. TechCrunch (2023). From July 6, 2024, source https://techcrunch.com/2023/06/22/get-a-clue-says-panel-about-generative-ai-its-being-deployed-as-surveillance-devices/. Accessed July 2024
67. Martínez, V.: Advierten por deepfakes para cometer fraudes. Reforma.com (2024). https://www.reforma.com/advierten-por-deepfakes-para-cometer-fraudes/ar2767208. Accessed July 2024
68. Ng, A.: https://twitter.com/AndrewYNg/status/1667920020587020290?s=20. Twitter. https://twitter.com/AndrewYNg/status/1667920020587020290?s=20 (2023). Accessed July 2024
69. Perrigo, B.: Exclusive: OpenAI lobbied E.U. to water down AI regulation. Time (2023). https://time.com/6288245/openai-eu-lobbying-ai-act/. Accessed July 2024
70. Rinaldi, L.: Rage against the machine [Interview with G. Hinton]. Toronto Life (2023). https://torontolife.com/deep-dives/geoffrey-hinton-sounding-alarm-artificial-intelligence/. Accessed July 2024
71. Roland, M.-C.: Claude 3 Opus has stunned AI researchers with its intellect and 'self-awareness'—does this mean it can think for itself? Livescience.Com (2024). https://www.livescience.com/technology/artificial-intelligence/anthropic-claude-3-opus-stunned-ai-researchers-self-awareness-does-this-mean-it-can-think-for-itself. Accessed July 2024
72. Shaban, H.: Elon Musk talks singularity with 'Rick and Morty' twitter account. Business insider (2017). https://www.businessinsider.com/elon-musk-rick-and-morty-singularity-2017-10. Accessed July 2024
73. Taneja, H.: The era of "move fast and break things" is over. Harvard business review (2019). https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over. Accessed July 2024
74. Warner Bros.: Picture presents an annapurna Pictures production; produced by Megan Ellison, Spike Jonze, Vincent Landay; written and directed by Spike Jonze. Her. Burbank, CA : distributed by Warner Home Video (2014)
75. Wiggers, K. Image-generating AI can copy and paste from training data, raising IP concerns. TechCrunch (2022). https://techcrunch.com/2022/12/13/image-generating-ai-can-copy-and-paste-from-training-data-raising-ip-concerns/. Accessed July 2024
76. Yudkowsky, E.: Pausing AI developments isn't enough. We need to shut it all down. TIME (2023). https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/. Accessed July 2024
77. Zakrzewski, C., Lima-Strong, C., Oremus, W.: CEO behind ChatGPT warns congress AI could cause 'harm to the world.' Washington Post (2023). https://www.washingtonpost.com/technology/2023/05/16/sam-altman-open-ai-congress-hearing/. Accessed July 2024