AI and Ethics
https://doi.org/10.1007/s43681-024-00600-9

ORIGINAL RESEARCH

Are the robots taking over? On AI and perceived existential risk

Airlie Hilliard[1,2] · Emre Kazim[1] · Stephan Ledain[3]

Received: 9 March 2024 / Accepted: 1 October 2024
© The Author(s) 2024

Abstract

Artificial intelligence (AI) is increasingly infiltrating our lives, and a large proportion of the population use the technology whether they know it or not. While AI can offer significant transformative benefits, this is only true if it is used in a safe and responsible way with the right guardrails. Indeed, there have been several instances of harm resulting from the use of AI without the appropriate safeguards in place. As such, it is unsurprising that there are mixed views of AI in society, where the negative view can in fact manifest as a dystopian view of "robots taking over". In this paper, we explore these positive and negative views of AI and the factors driving such perceptions. We propose that negative perceptions of AI often concern job displacement, bias and fairness, and misalignment with human values, while positive perceptions typically focus on specific applications and benefits of AI, such as in scientific research, healthcare, and education. Moreover, we posit that the types of perceptions one has about AI are driven by their proximity to AI, whether general or specific applications of AI are being considered, knowledge of AI, and how it is framed in the media. We end with a framework for reducing threat perceptions of AI, such that the technology can be embraced more confidently in tandem with risk management practices.

Keywords Artificial intelligence · Bias · Value alignment · Automated decision systems

Playful, fun -
The moment is alive I often catch her laughing at me
Though I don't know why
But I feel the joy - her joy
It spills from a place of ecstasy -
In those moments
She gives herself away
And so I simply embrace it
Sinking into her eyes that
Set like jewels upon
The throne of her perked cheeks
Her smile envelops me
Snatching my breath in an absolute
It's garden of roses and tulips
In full bloom - soaked
In a Sun that never sleeps
-Tahsin Beyazyurek

1 Introduction

Readers might recall that in summer of 2004, movie theatres buzzed around Will Smith's war against the robot rebellion in "I, Robot." His character, Del Spooner, passionately warned viewers of the dangers of relying on machines for crucial ethical decisions. This sentiment, echoed in similar movies before and after I, Robot and shared across water coolers and classrooms for decades (e.g., [1]), has woven itself into a collective narrative about artificial intelligence (AI) among the general public. Accordingly, this has fuelled mass hysteria around robot-induced doomsday, with almost 34,000 people signing an open letter by the Future of Life Institute [2] calling for a six-month pause on the training of AI systems more powerful than GPT-4. As a result, a fear-laden shadow has been cast over the public's collective willingness to understand and participate with AI systems. Yet, rarely is there a counterpunch to this dystopian view,

✉ Airlie Hilliard
airlie.hilliard@holisticai.com

[1] Holistic AI, 18 Soho Square, London W1D 3QH, UK

[2] Institute of Management Studies, Goldsmiths, University of London, New Cross, London SE14 6NW, UK

[3] University College London, Gower Street,
London WC1E 6BT, UK

Published online: 15 November 2024

or a vision of a more mindful, harmonious adoption of AI. But perhaps the true danger, the one yet to be captured on screen, lies not in robot overlords, but in our disengagement from shaping the future of this powerful technology.

A relatively young technology, AI has come a long way since the concept emerged in the mid twentieth century, first being established as an academic discipline by Alan Turing in 1950 [3] with the publication of his seminal article questioning whether machines can think and proposing that computers could become apt at games such as chess. Indeed, some of the earliest applications of AI focused on niche if–then scenarios, with IBM's Deep Blue chess-playing program beating the world chess champion in 1997 [4]. Demonstrating the significant advancements in the capabilities of AI, Deep Mind's AlphaGo beat the Go world champion in 2015, which required much more sophisticated programming to navigate the more complex rules and gameplay compared to the chess program [4].

Contemporary artificial intelligence has now diffused beyond isolated research contexts and games into impactful real-world implementations that can affect daily life and carry significant implications for everyday citizens. For example, AI can be used in the healthcare context to recognise tumours with high diagnostic accuracy to reduce the burden on doctors [5], support agriculture through the early detection of plant disease [6], and to streamline customer support and elicit more consumer feedback [7]. The realisation of these benefits has seen a dramatic uptake in the everyday use of AI, with 35% of companies around the world using AI already and a further 42% exploring its use [8], and the AI market set to breach the $500 billion mark by 2025 [9].

However, with novel technology comes novel risks; AI has already been involved in several high-profile scandals such as Amazon's retired resume screening tool that was biased against females [10], and Northpointe's COMPAS tool to predict recidivism that was biased against black defendants [11]. More recently, there has also been an emergence of lawsuits due to the alleged unfair use of AI, including against Lemonade Inc for the use of biometric data [12], State Farm for allegedly having more difficult claims processes for black policyholders, [13] and Applicant Tracking System provider Workday for allegedly biased algorithms [14]. Indeed, while AI can undoubtably offer unmatchable capabilities, arguably, much of the attention that AI and other algorithmic systems receive is negative, fuelling concerns about the harms that AI can bring when left unchecked.

Alarmingly, almost 75% of those using AI have reportedly not taken steps to ensure trustworthy and responsible AI, almost 70% do not monitor model performance, and only around 40% have taken steps to increase the explainability of their AI systems [8]. This inaction is a concern shared by the public, industry, policymakers, and academics alike, with new fields such as ethical AI, trustworthy AI, and responsible AI [15–17] emerging to support the implementation of guardrails to prevent avoidable harms. Several laws have also been proposed around the world to regulate AI and algorithmic systems, particularly in the US and EU. These laws impose varying requirements, from bias audits and notification for employers or employment agencies using automated employment decision tools in New York City under Local Law 144 [18] to sweeping requirements touching on issues such as transparency, data quality, risk management, and record keeping across sectors under the EU AI Act [19]. While these efforts are certainly a step in the right direction, public concerns about the potential risks posed by AI are still lingering, with the technology met with scepticism about its fairness, usefulness, and risks [20].

However, despite depictions of robots taking over the world in the media, as we will explore, superforecasters and AI experts only predict there to be 0.38% and 3% chance, respectively, of AI causing the extinction of humanity, so the threat of AI completely taking over is low [21]. If this is true, then what is driving public concerns about 'robots taking over'? In this paper, we address this question by first disentangling AI and robots before exploring the positive and negative perceptions of AI taking over, particularly focusing on job displacement caused by AI and automation and the AI alignment problem. We then explore what is fuelling concerns about the existential risk of AI before proposing what purpose such beliefs serve and how AI can be adopted with greater confidence. In particular, we posit that one of the greatest drivers of concerns about AI and existential risk is a lack of education on AI, its capabilities and limitations. Accordingly, we suggest some actions that can be taken by multiple AI stakeholders to reduce the existential threat of AI and shape the technology to serve humanity positively and in a way that fosters innovation. In doing so, we acknowledge the complex landscape of AI through a common understanding of what it is and is not capable of in hopes of rooting one's pessimism (and optimism) in real and existing applications of these tools. This in turn aims to debunk myths about AI and offers ideas about how to address existing ethical concerns and potential points of misalignment. A conceptual overview of the paper can be seen in Fig. 1.

The intention of this paper is to extend the conversation around the risks of AI, public perceptions of these risks, and how concerns might be managed to increase the confidence in and safety of AI adoption among the public, or non-AI experts. Specifically, through an examination of public perceptions of AI and their origins, our goal with this paper is to provide a framework for AI alignment and risk reduction, including AI assurance and monitoring, education and

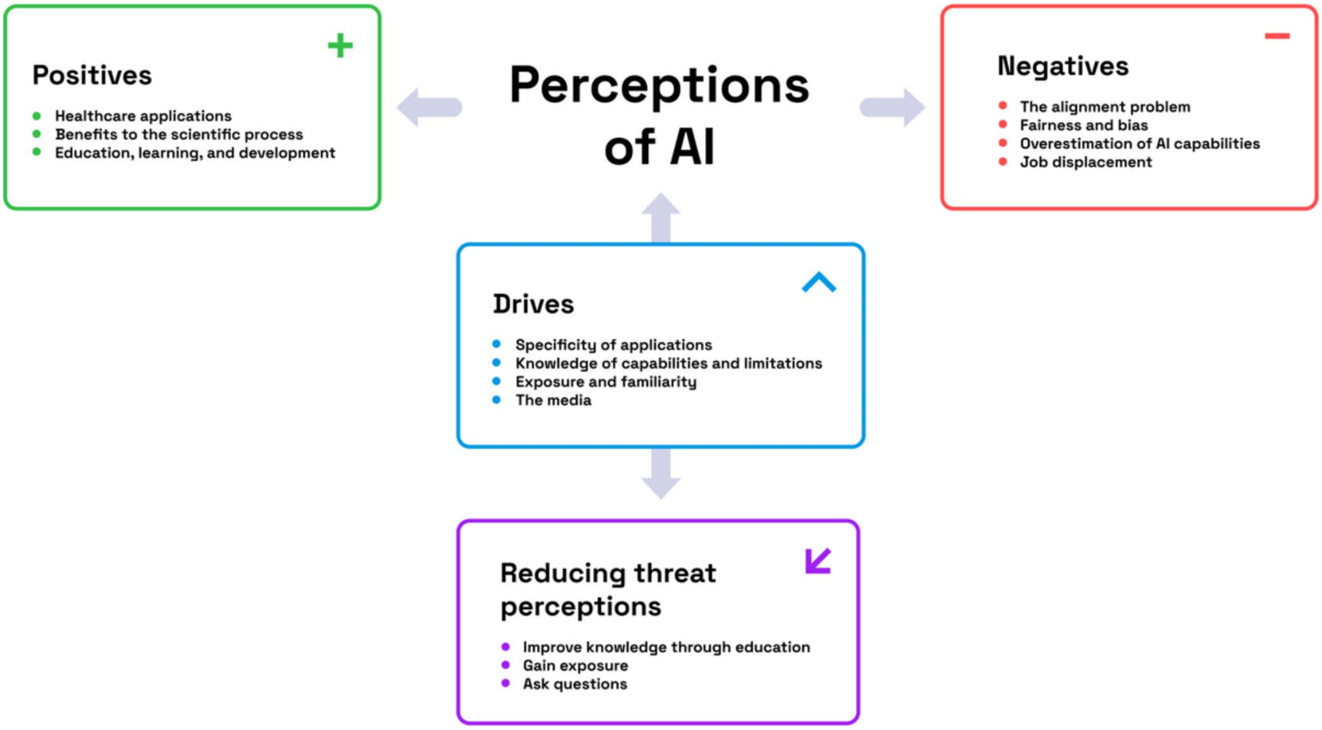
Positives
+
Healthcare applications
Benefits to the scientific process
Education, learning, and development
Perceptions of AI
Negatives
−
The alignment problem
Fairness and bias
Overestimation of AI capabilities
Job displacement
Drives
Specificity of applications
Knowledge of capabilities and limitations
Exposure and familiarity
The media
Reducing threat perceptions
Improve knowledge through education
Gain exposure
Ask questions

Fig. 1 Visual overview of the paper structure and key arguments

awareness, and civic engagement in order to create a greater alignment with AI to reduce its threat and create progressive pathway forward. This serves as a way to understand how, in this time, individuals can actively participate in shaping and aligning their personal aspirations with the capabilities of these tools. Through informed participation, we can shape AI to serve humanity positively, fostering innovation, sound policies, and meaningful applications that benefit us all.

2 The misconception about AI equalling robot

Before examining the factors influencing perceptions of robots taking over, it is important to understand how AI can be conceptualised by the public and acknowledge misconceptions about AI being analogous to robots. Indeed, for some laypeople, AI and robots are commonly equated [22] and seen as one and the same or as sharing similar risks and threats to humans and society (e.g., [23]). However, for our discussion, it is important to understand the distinction between the two technologies, although this is not a simple task; there is a lack of alignment on exactly what is meant by artificial intelligence, with different entities proposing vastly different definitions. One of the most converged upon definitions of AI is the Organisation for Economic Co-operation and Development's [24], where an artificial intelligence system is defined as:

> *"a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy"*

This definition has influenced a number of policymakers seeking to regulate artificial intelligence, with the definition provided by the European Union's AI Act [25] pivoting from a lengthier definition of AI to a more succinct one that heavily draws upon the OECD definition, although the AI Act definition asserts that AI systems can vary in their adaptiveness after deployment. Other proposed laws such as the now-dead California Assembly Bill 331 [26], which sought to regulate a range of AI systems and impose requirements for developers and deployers, and the AI Risk Management Framework published by The National Institute of Standards and Technology [27] also used a similar definition to the OECD definition.

Among other definitions of AI that do not draw so heavily on the OECD definition, however, are some commonalities wherein four key elements are typically captured: automation or autonomy, human influence in providing inputs and defining objectives, various outputs, and the technology underpinning the system. For example, Canada's Artificial Intelligence and Data Act [28] defines AI as:

> *"a technological system that, autonomously or partly autonomously, processes data related to human activities through the use of a genetic algorithm, a neural network, machine learning or another technique in order to generate content or make decisions, recommendations or predictions."*

Others propose a more unique definition of AI, with the Council of Europe asserting that AI, "brings together sciences, theories and techniques (including mathematical logic, statistics, probabilities, computational neurobiology and computer science) and whose goal is to achieve the imitation by a machine of the cognitive abilities of a human being." [29]. Similarly, Connecticut Senate Bill 1103 [30] provides a particularly lengthy definition of AI, in which it claims that AI:

> *"Is designed to (I) think or act like a human, including, but not limited to, a cognitive architecture or neural network, or (II) act rationally, including, but not limited to, an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communication, decision-making or action, or (B) a set of techniques, including, but not limited to, machine learning, that is designed to approximate a cognitive task"*

These definitions converge more with the concept of artificial general intelligence (AGI), where autonomous systems and machines would be capable of human-like intelligence including abstract learning [31]. Although AGI is still hypothetical, there are signs that the technology could be developed in the future, with the AI-driven humanoid robot Sophia, developed by Hanson Robotics, obtaining citizenship in Saudi Arabia and being named Innovation Ambassador for the United Nations Development Programme [32]. However, it is also important to acknowledge that not all robots are AI-driven, and AI and robots are not one and the same. Indeed, according to the European Parliament, a *smart* robot can acquire autonomy through analysing data collected sensors or otherwise from its environment, carry out self-learning from experience, adapt its behaviour to its environment, and has at least minor physical support, all while having an absence of biological life [33]– sharing some notable similarities to the definition of AI.

On the other hand, physical robots can be defined as a reprogrammable, multifunctional machines that are designed to move specified features of the environment through variable programmed motions [34]. As such, robots can be thought of being on a spectrum, from machine-based robots explicitly programmed to perform a specific task to smart robots that have greater autonomy and that are AI-driven. In short, robots can be powered by AI, such as in the case of Sophia, but the two can also exist in the absence of each other. Therefore, robots and AI are distinct technologies that can be integrated to form intelligent machines that can interact with and make inferences about their environment.

3 What are the perceptions of AI taking over?

While the common understanding of "robots taking over" will have negative implications, this may not be necessarily true for all people nor in the same way. Indeed, such perceptions can be influenced by individual situational and contextual factors including one's technical expertise and education, exposure to and familiarity with AI, and subgroup membership. Identifying and understanding both the positive and negative perceptions of "robots taking over" is an important step in understanding what drives them. As such, in this section, we discuss some of the key positive and negative perceptions of robots (read AI) taking over that are widespread in society.

3.1 Negative perceptions

3.1.1 The alignment problem

A focal point of fear and apprehension around the emergence of AI is the Alignment Problem. Simply defined, the Alignment Problem involves ensuring artificial intelligence systems continue to behave in ways that are beneficial to humans as the capability of these systems increases. The risk of these tools growing away from human needs is at its core an issue of control [35]. For example, an intelligent AI, could pose a threat if it operates solely on misaligned goals and incentives, or further, ones that are not equitable and inclusive of broad human desires [36]. An AI-powered hiring or criminal justice system that is not properly aligned may inadvertently discriminate against certain demographics, leading to further entrenching biased selection processes and perpetuating systemic inequities. That is not to underappreciate that human values and goals are as complex as they are varied, context-dependent, and challenging to define in a system, making them difficult to program [37]. Although not quite the army of killer robots we may fear, misaligned AI is the first step on the path to the doomsday scenario of humanity's loss of agency and inability to determine future outcomes.

3.1.2 *Bias* and fairness

Related, fear of AI taking over stems from issues of bias and unfairness that can arise in multi-faceted ways every stage of the algorithmic lifecycle. At a non-intuitive level, algorithmic systems (such as social media timelines) synthesize information that aim to remove randomness and serendipity. They continually expose people to only familiar and logical pathways [38]. As a result, important discoveries or opportunities that depend on randomness or flexible thinking could be missed. At a more intuitive level, biases may exist in how data is collected to train these models, with certain populations overrepresented in ways that skew decision outcomes. Machine learning models also learn from historical data [39]. If that data contains biases, the model will reproduce those biases at a large scale and amplify and entrench harmful outcomes. Notably, Google Photos' image recognition algorithm incorrectly labelled some Black people as gorillas, and Amazon developed a hiring algorithm in 2018 that discriminated against women [10, 40]. Next, as these algorithms begin to interact with each other, unforeseen negative outcomes that a difficult to pinpoint or reverse may emerge. With AI increasingly being used in critical applications such as healthcare, employment, law enforcement, and education, it is understandable that such high-profile cases have significantly shaped public perceptions of AI and have resulted in apprehension about interacting with the tools for fear of being disadvantaged, particularly among underrepresented groups.

However, many of these biases reflect systematic societal biases that infiltrate systems in their design, development, and training [41]. While AI can result in more widespread harm than human decisions due to its scalability, it is important to recognise that these phenomena are not novel to AI. Further, despite seemingly widespread concerns about the potential for AI to be biased, interestingly, there is evidence to suggest that there is less moral outrage associated with algorithmic discrimination compared to human-driven discrimination since AI does not have the sentience to purposefully discriminate. Instead, discriminatory outcomes are a byproduct of existing human biases, although there is seemingly a lack of blame on human creators of discriminatory AI tools [42]. However, increasing instances of social disparity may emerge because of AI, not necessarily within the tools themselves. For example, wealthier companies in wealthier economies will have the resources to build, leverage, and reap the benefits of AI. Related, the issues AI is purposed to address will reflect those of that class that can afford to develop AI [43].

Moreover, it is important to remember that group differences in predictions from AI systems is not always indicative of bias. Indeed, the power of algorithms and AI is in the ability to recognise patterns in data, particularly patterns that may not be intuitive or recognisable by humans [44, 45]. As such, AI may conversely be able to uncover genuine underlying differences between groups and individuals that can have value for predictions. Therefore, mitigating all subgroup differences may result in homogeneity of predictions and reduce their utility if the underlying reason for subgroup differences is not investigated. To give an applied example, there is a century of research into the value of individual differences in the context of job recruitment, where differences in personality and cognitive ability, for example, can be useful for predicting job performance [46–48]. If these differences in predictions are due to genuine differences in ability and not due to biases, then adjusting the model to result in more even outcomes may reduce both the validity and utility [49] if this limits the ability to set individuals apart and identify top performers. This is all to say that fear of impending doom can look like larger-scale, more insidious version of the harmful biases already found in existing social institutions where decisions about meaningful life outcomes exist beyond one’s control.

3.1.3 Overestimation of AI capabilities

The above fears surrounding AI are often fuelled by a sensationalized narrative that exaggerates its capabilities [50]. This misrepresentation amplifies concerns about AI dominance and our ability to manage algorithmic systems. Historically, when it comes to technological emergence, humans tend to overestimate the potential and their societal impact. A prime example is the prediction of universal availability of fully autonomous vehicles by 2020, which in 2024 remains far from reality [51]. This can be driven by what Gartner has coined as “the hype cycle” of innovative technologies, a cycle of which AI has not been exempted [52]. This cycle is reflected in the media, where coverage of AI advancements so far has tended towards hyperbole and, in the more conscious publication, speculation. For example, news outlets have exaggerated the imminence of artificial general intelligence (AGI), although there has been little no evidence of such breakthroughs, focusing on the negative outcomes and risks. This emphasis on risks, without adequately contextualizing limitations, has distorted public understanding of AI's potential [53].

This can also be compounded by what cognitive scientists have termed the Dunning-Kruger effect, where individuals overestimate their expertise in complex domains [54]. The technical complexity and rapid advancement of AI create a knowledge gap between experts and the public. As such, thought leaders and prominent figures can tend to overestimate their understanding of the technology, creating cycles of misinformation. What’s more, research suggests that limited knowledge in a subject area increases susceptibility to false information [55], and while we overestimate our ability to predict AI capability, we underestimate the complexity of human intelligence. This in turn, further contributes to the overestimation of AI capabilities. Indeed, human intelligence encompasses a range of multifaceted abilities that operate in ways not yet fully understood. The human brain's ability to integrate and actualize these modes of intelligence such as empathy and cognitive reasoning, in a broad range of contexts, is still unmatched [56]. This underestimation fuels a misconception around AI capabilities but moreso the risks associate with them as they pertain to our ability to be replaced or deceived by them.

3.1.4 Job displacement

Building on this, it is estimated that up to 800 million jobs might be replaced by automation by 2030, with up to 375 million needing to switch occupation as their role is increasingly replaced with technology [57]. This will likely affect certain industries more than others, with roles that are heavily process-based such as machine operators and assemblers facing a much higher risk of automation at 60% than professionals, senior managers, and officials, who only face around a 10% risk of automation [58]. This is a trend that has been observed a number of times before; with the advent of increasingly advanced technology and each industrial revolution, of which AI is part of the fourth industrial revolution, jobs created by the previous revolution or earlier can get displaced [59]. Nevertheless, AI and automation do not appear to pose as big a threat as can be catastrophised; despite the number of companies reporting using AI increasing four points from 31% in 2021 to 35% in 2022 [8], the global unemployment rate fell between 2021 and 2022 [60], indicating that overall, AI and robotic automation are not posing a significant threat to jobs that is leading to increased unemployment.

Indeed, changes in the way that we work can provide novel opportunities for innovation, create new roles, or lead to the upskilling of employees. In addition to the increased demand for roles related the development, maintenance, and sales of AI systems, such as machine learning specialists and data scientists, the widespread adoption of AI has seen the creation of new roles such as AI auditors and AI policy experts as legislative efforts proliferate [61]. Further, using AI to automate repetitive tasks can increase human capacity for more creative tasks. A real-life example of this is Ikea customer service workers, who were upskilled and trained in interior design when their jobs were displaced by an AI chatbot [62].

Moreover, recent research into the benefits of generative AI in the workplace, specifically ChatGPT, indicates that the use of AI in the workplace can increase output quality by almost 20% and reduce time taken to complete tasks by 40% [63]. However, it is estimated that clerical work is the most exposed to the effects of generative AI while other tasks have much less exposure with an estimated 1–4% of other tasks having a high level of AI exposure compared to 24% of clerical tasks. As such, in many of the cases, AI, particularly generative AI, typically is used to support or augment human activities, rather than being used for full automation or replacing humans. Greater education on the tools as well as AI alignment can, therefore, help individuals to harness their potential and make themselves more efficient, reducing their risk of being displaced by the tools altogether.

3.2 Positive perceptions

It may be cliché to state that AI stands as a rapidly evolving domain of innovation that carries its own nuances and is incredibly difficult to compare to any other. However, like any novel technology, public discourse around AI tends to split between utopian and dystopian perspectives. On one side, discussion focuses on prospective benefits and opportunities enabled by AI. Counterbalancing this, another strain weighs potential risks and negatives if development outpaces ethical safeguards and alignment to the many varied human needs and requirements. With some objective distance, a balanced view acknowledges the nuanced space between these poles and arrives at what may prove to be the most accurate. By considering both upsides and downsides of the adoption of AI, we gain a fuller understanding of AI's multifaceted impacts across contexts. As such, in the following sections, we explore some of the specific applications of AI that can be perceived to bring widespread benefits to the general public if its use is scaled up: healthcare, scientific discovery, and education.

3.2.1 Healthcare

Within healthcare, Artificial Intelligence holds immense potential to unlock breakthroughs and democratize access to life-saving treatments if adopted at a wide scale. For example, AI systems can rapidly sequence human DNA, analyze complex molecular interactions, and identify promising new drug targets—accelerating the pace of discovery exponentially [64]. For clinicians, AI can ingest vast medical databases and patient histories to assist in early disease diagnosis, personalized care plans, and predictive analytics to determine best interventions based on data-validated outcomes research [65]. Robotic surgical aids supervised by AI can conduct tireless microsurgeries with enhanced precision beyond human limitations [66]. Telehealth powered by natural language processing and computer vision lets doctors diagnose and advise patients in remote areas lacking on-site medical infrastructure [67]. Some of this optimism has been met with caution. A systematic review of the role of AI in healthcare practices, highlights the potential danger of AI reducing the confidence and even the efficacy of medical professional judgement [68]. Then there's evidence to suggest that patient perceptions of healthcare service differ in the degree and methodology of this care. For example, Palmisciano et al.'s study [69] found that some patients had varied comfort levels depending on the context and use of AI applications and expressed a preference for having some degree of human support in the process.

With pragmatic optimism, we can leverage these technologies to democratize healthcare access for underserved communities, catch illnesses earlier, and provide the best evidence-based treatments for all. Indeed, while AI's capabilities are not yet sophisticated enough for it to fully take over and replace healthcare professionals, it can be applied strategically to support them and improve patient care, with robots (AI) taking over in this sense typically being perceived positively, providing that it is used with expert supervision and the appropriate safeguards in place.

3.2.2 Scientific process

Further and more broadly, the application of AI in practices such as scientific discovery has received much acclaim through its ability to amass and analyse larger amounts of data and uncover more specific patterns. For example, in the field of astrology, AI can expand our understanding of the universe by looking at data gathered from telescopes and infer the location of celestial bodies, data that humans would otherwise find too large to analyse alone [68]. Beyond analysis, AI can support human efforts in asking more sophisticated questions of the universe. With its ability to generate more sophisticated statistical models, the speed at which hypothesis testing and building simulations occur through the use of AI can support efforts to reveal what questions are even possible to ask [70].

What's more, AI's augmenting of scientific research is not confined to "hard" sciences but expands into the more nuanced social sciences that have historically been seen as"messy". AI's ability to analyse large amounts of behavioural and interview data for example, similarly, possesses the ability to uncover subtler trends that would be difficult for humans to detect alone [72]. This allows researchers to generate new hypotheses about socially occurring phenomena, and in turn, empowers researchers to investigate more complex social phenomena and expands our ability to understand individuals and groups [73]. A meaningful limitation of these capabilities is that not all patterns are practically or clinically significant. Some discoveries can lead to misleading conclusions that adversely affect life outcomes. These algorithms may not account for all the complex factors that influence behaviour and require additional interpretation by researchers who can consider the context and application of the findings (Constantino, Schlüter, Weber, & Wijermans, 2021).

Uncovering new insights need not stop at any scientific discipline, but AI could reveal cross-disciplinary insight and advance our ability to understand the interconnectivity of observable phenomena. It can be said that much of the research process is in tedious routine tasks such as data labeling and annotation, and through an enhanced automation capability, researcher time can be freed up to address more complex problems that yield higher impact [74]. As AI systems continue to evolve and become more sophisticated, they are likely to play an increasingly important role in scientific research. Indeed, DeepMind's AlphaFold is already paving the way for the acceleration of scientific research with a public database of predicted protein structures derived using complex AI algorithms [75]. Yet, with all this capability, there is justification to focusing on both the"how" and the"why" underpinning AI models. For example, explainability ("how ") of AI models that conduct these studies will be critical. A recent article explores the many risks present in applying AI discoveries without understanding of the underlying decision-making architecture in a host of domains including medical and biomedical, healthcare, finance, law, cybersecurity, education and training, and civil engineering [76].Generating insight without sound theory ("why") may lead researchers to struggle to root these findings in real world application, a critical step in enhancing the positive valence of AI-based research [77].

3.2.3 Education, learning and development

The use of algorithms and AI in education can be controversial and lead to mixed perceptions on its efficacy and appropriateness. However, we argue that perceptions and indeed applications can be more positive when AI is applied for specific tasks to enhance education and skills development in various ways. For instance, AI-powered adaptive learning systems can personalize learning experiences for students, tailoring the content and pace to their individual needs, abilities, and learning styles [78]. This can lead to more effective learning outcomes and increased student engagement and ultimately have students feel all around more satisfied and equipped through the learning process. AI can also assist teachers in grading and providing specific feedback, freeing up time for more hands-on, human interaction with students such as interpreting the feedback. Additionally, AI can help identify skill gaps and recommend targeted training and developmental programs for workers, empowering them to upskill and reskill in an ever-changing job market [79]. AI-powered virtual teaching assistants can also facilitate collaborative learning and simulate real-world environments, enabling students to practice skills in a more immersive and realistic way [80]. By augmenting human capabilities, AI can help education and training institutions keep pace with the rapidly evolving demands of advancements in the digital age and be perceived as a positive resource.

4 Where do these perceptions come from?

It is clear that depending on the context in which AI is thought of can result in different perceptions of AI, where individuals may have a mixture of both positive and negative views of the technology overall. In order to understand which side may be more prominent in individuals, it is important to understand the factors that drive such perceptions of AI. As such, in this section, we propose some key shared drivers of AI perceptions that can provide insights into sources of existential concerns about AI.

Firstly, the exploration of the positive and negative perceptions of AI above reveals an interesting insight; many of the negative perceptions of AI and associated existential catastrophising are associated with general purpose AI systems [81] while more positive perceptions stem from AI systems designed for a particular application. For example, although there are justified concerns about the potential for harmful and hallucinatory outputs of generative AI– which is a general purpose tool until it is refined for a particular application– specialist AI tools are widely accepted and used in other contexts such as healthcare, where 68% of physicians are excited about the role of AI in healthcare and 31% have used it to support their patient care [82].

This emergent dichotomy between public reactions can be better understood through the lense of cognitive psychology. Cognitive biases arise in the face of novel information as individuals often take mental shortcuts known as heuristics to draw meaning, a well-documented phenomenon [83, 84]. For example, the tendency to avoid uncertainty and ambiguity can play a role [85]. General purpose AI systems inherently introduce uncertainty because they lack clearly defined use cases. The lack of specificity can introduce cognitive biases through increased speculation and doubt of who is accountable for negative consequences, activating what is known as the attribution of responsibility problem [86]. Whereas specific use cases, especially those with more intuitive applications towards public good, can potentially address anxiety through added clarity.

Framing effects, as posited by Tversky and Kahneman [87], can shape how one makes decisions and forms ideas around risk. In reflection of how AI has been introduced to public discourse, specific use cases are typically framed within the context of social goods. Further, specific applications are often framed within familiar technological contexts allowing individuals to draw on readily available examples, a concept known as the availability bias [88]. Use cases seen in healthcare and education for example, are positioned within the context of benevolent institutions who have successfully leveraged technology creating a positive association. While framing effects largely explain the positive perceptions of specific AI applications, they also shed light on why general-purpose AI often evokes more negative reactions.

The widespread potential harm and negative press of general-purpose AI has created a general valence of mistrust [89]. This can be seen through the lens affect heuristic biases where is feeling dread or enthusiasm impacts how one feels towards risk (more enthusiasm equals less risk). While this is not a phenomenon that has been widely studied, we propose that this could be driven by how widespread potential harm resulting from these systems could be. Indeed, what is often seen as an exciting and unforeseen feature of generalized models, the ability to apply and generalize knowledge to unforeseen situations can lead to unintended and unpredictable outputs, further fuelling mistrust. Therefore, the harms from the misuse of general-purpose tools can be more widespread and affect larger groups within society than those designed for a specific task, which could fuel greater distrust in the tool [90]. For example, a single output of a credit scoring algorithm will have important implications for a single individual at a time, whereas a single deepfake may circulate around the internet and go viral, therefore affecting potentially millions of individuals. This is a particular concern if maliciously to spread disinformation around elections, for example, since voting behaviour may be influenced by unreliable sources, which can in turn reduce trust in news outlets [91]. As such, there has recently been a wave of activity, particularly in the US, targeting deepfakes in elections [92], while AI in credit scoring is being prioritised to a lesser extent.

There are also important considerations relating to the training of AI that must be taken into account given that many of these models were trained based on arguably outdated datasets. In the case of ChatGPT, the GPT 3.5 model was trained on data scraped up until January 2022 [93], meaning that any new information that has become available since then is not reflected in the training data of the model. Therefore, any outputs of the model are based on data that may not reflect current knowledge. It is this lack of understanding about the models and their limitations that could pose the biggest risk from the use of these tools wherein an overreliance on the tools could develop due to their almost limitless applications [94], leading to them being used as shortcuts without their outputs being verified. Consequently, it is essential that the tools are only to be used by individuals for tasks that they have the relevant expertise for to verify their outputs to prevent avoidable harms. As such, a lack of knowledge of how general-purpose, publicly accessible AI tools work and their limitations or a lack of explanation of such can lead to negative perceptions of the utility of AI among the public if outputs are not as expected [95].

Secondly, exposure to AI tools can shape perceptions, where the more exposed individuals are, the more accepting of and optimistic about the tool they are. For example, almost double the number of regular users of generative AI are optimistic about the technology than those who do not use generative AI [96], where frequency of use is positively related to likelihood of using the tool in the future [97]. This sentiment is also echoed outside of generative AI, in applications such as recruitment [98]. This concurs with the fact that the perceived trustworthiness of AI systems has increased between 2020 and 2022 [99], a trend that is likely to continue. We posit that this is because as users interact with the tools more, the more familiar they become with the tool's benefits, limitations, and potential risks. Indeed, those that receive training in AI are more confident in their use of the tool in the future [100, 101], meaning that a lack of AI literacy could fuel existential thoughts about AI taking over. As such, those that educate themselves more on how AI systems are trained will be aware of the limitations with their training data and therefore avoid using them as an information source for current events, likely leading to more positive perceptions. In other words, individuals that are familiar with and educated on AI can have more positive perceptions of the technology than those who are not since they can set realistic expectations for their interactions and will be more aware of the maximum capabilities of AI such that they can complement the technology with their own abilities to complete tasks. As such, this could reduce the perceived threat of AI-driven job displacement, for example, if an individual is able to use the technology to complement their own efforts to increase productivity and performance.

Thirdly, the media significantly shapes views of AI [102] and can subsequently fuel perceptions of robots taking over. Indeed, reporting in this space by the media focuses on harms and lawsuits [103], with AI benefits and successes less reported. Likewise, there are a number of repositories that track AI harms across applications (e.g., OECD AI Incident Monitor, AI Incident Database; AIAAIC Repository; Holistic AI Tracker), but there is a lack of an equivalent for positive outcomes. While there must be awareness of such issues to promote responsible AI and highlight the need for AI risk management, without the right balance with positive use cases, this is likely to fuel existential thoughts about AI and decrease trust. Moreover, pop culture can add to this. Although there are cases where AI is seen as a force for good, for example, in Wall-E and Robot & Frank, the vast majority of cinematic depictions of AI has been in terms of a threat and/or dystopia. Some examples: Metropolis (1927); 2001: A Space Odyssey (1968); Westworld (1973); Star Trek: The Motion Picture (1979); Alien (1979); Blade Runner (1982); Tron (1982); The Terminator (1984); Short Circuit (1986); RoboCop (1987); Terminator 2: Judgment Day (1991); The Matrix (1999); Bicentennial Man (1999); A.I. Artificial Intelligence (2001); I, Robot (2004); Prometheus (2012); Ex Machina (2015). As such, depictions AI as an existential threat are reinforced over time and across medium, which can cause uncertainty and information bias [104], thereby making individuals more prone to confirmation bias of negative perceptions of AI.

As such, an individual's perceptions of AI are shaped by a variety of forces, and the sum of these forces determines whether an individual has existential concerns about AI taking over. Indeed, individuals low in AI exposure with limited understanding of its limitations that are exposed to negative media content are more likely to perceive AI as an existential threat, while individuals who regularly interact with AI and are educated on how tools work and their limitations may be more resilient to viewing AI as an existential threat even when exposed to negative media stories.

5 How can (the perception of) existential risk be reduced?

One of the major factors driving negative perceptions of the existential risk of AI is a lack of knowledge of or indeed education on the technology [96, 100, 101]. This lack of knowledge and education can result in actual risks of AI being realised if those deploying and developing the systems lack the expertise needed for technical and governance approaches to manage AI risks. As such, there is a need for more education and outreach to improve general AI literacy and address the misconceptions that fuel fears about "robots taking over", as well as efforts to reduce instances of harm that can further fuel negative perceptions. This requires a collaborative effort from multiple stakeholders, including policymakers, AI developers and deployers, and users in order to effectively reduce existential perceptions of AI and ensure the technology is used in a way that is aligned with human values and can support safe innovation. As such, in this section, we suggest some actions that can be taken by these different stakeholders to improve AI perceptions, which we summarise in Table 1.

Table 1 Recommended actions to reduce the actual and perceptual threat of AI across different stakeholders

Stakeholder	Recommended actions
Developers and deployers	Greater transparency and explainability, such as through transparency statements Ethical self-governance through policies, procedures, and practices Monitoring throughout the system lifecycle Risk mitigation
Policymakers and judges	Create and enforce AI-specific laws Hold those that do not comply with legal requirements accountable Improve their own AI knowledge through training and education
Individuals	Use AI in a way that is consistent with system instructions and limitations Seek out and identify formal or informal AI educational opportunities Actively seek out balanced views on AI
Academics	Research the efficacy of self-governance frameworks and laws on improving trust and other outcomes Deliver training sessions

As the entities creating and making AI systems available on the market, developers and deployers play a key role in shaping public perceptions of AI through their own actions. Indeed, a significant driver of distrust in AI is a lack of transparency [105, 106] and less than a third of the public can confidently explain how AI works [107]. Accordingly, greater transparency around AI increases perceived effectiveness of AI and promotes trust [108]. However, the language used in such transparency efforts must be carefully considered to avoid causing additional concerns about issues such as privacy [109, 110]. Nevertheless, providers of AI and algorithmic tools have already started to take steps toward this, with X (formerly Twitter) open-sourcing its recommendation algorithm and accompanying documentation [111] and video interview and algorithmic assessment provider HireVue issuing an explainability statement [112] to provide information about how their tools work. In the absence of codified requirements, industry self-governance can help to reduce promote trust as well as mitigate risk through well-defined processes, policies, and procedures that are centred around the ethical use of the technology [113–115], where AI governance can be applied throughout the entire lifecycle of the tool [116] in order to make sure that AI is aligned with human values. As such, identified risks such as a lack of fairness and bias can be mitigated before they cause harm, reducing the potential of AI becoming a threat and facilitating safe innovation.

Despite self-governance and internal policies offering a potential route to increasing trust in AI and reducing actual harm, policymakers and regulators also play a crucial role in codifying these best practices into law, and judges must hold those who do not comply accountable. Indeed, legislating AI could help to increase trust in the technology [117]. While there has already been some progress towards this in the employment tool domain for example, with New York City Local Law 144 requiring employers using automated tools to make employment decisions to acquire an independent, third-party audit since 5 July 2023, wider progress is lacking. For example, the EU AI Act is expected to become the global gold standard for AI legislation, but its most stringent obligations will not be enforceable until the middle of 2026. Nevertheless, many market players have already started their compliance journeys in anticipation for these rules, with over 550 entities having signed up for the AI Pact, a voluntary commitment to early compliance with the Act's requirements [118]. Moreover, it is important that policymakers are educated on the tools they are seeking to impose requirements on in order to ensure that they are actionable and will have the intended effect. As such, policymakers, enforcement bodies, and those working in the legal system should engage in additional training on AI if required. European judges have already made progress on this recently with a training course offered by UNESCO and European Judicial Training Network focusing on how AI can be leveraged in the judicial system [119].

In addition, users of AI systems can play their own role in increasing their own trust. Indeed, individuals who intend to use the technology should ensure that they have an adequate understanding of the technology they are going to engage with, including accessing any available explainability statements, reading instructions, and using systems in the way that is intended and in line with system limitations. This could help to address issues caused by an overestimation of the capabilities of AI, reduce overreliance, and support the use of AI in a more productive and supplementary way. Moreover, a balanced understanding of media stories can help individuals to act judiciously with respect to AI [120]. Doing so may require a conscious effort to resist confirmation bias and seek out more positive AI use cases and examples of risk mitigation from credible sources. To reduce the influence of misalignment on AI perceptions, individual should also evaluate their own unique stake in the evolution of AI while aligning it with their individual interests. They could also outline clear goals and actions for [112] both continuously expanding one's AI literacy but also contributing and shaping the conversation more broadly. For example, individuals could actively participate in discussions and provide commentary on industry forums and social media and engage in education opportunities where possible, as thousands of students have already [121].

This journey must also prioritize inclusivity and diversity. That also means AI literacy should not be limited to technical experts or those who use the technology for work or to optimize performance. A common language and understanding around AI is needed to avoid alienating those who are less familiar with the technology. As such, reducing the actual and perceptual threat of AI should be a joint endeavour whereby multiple and diverse stakeholders form ongoing collaborations Moreover, academic researchers can play a key role in this by researching the effectiveness of self-governance, legislation, and individual actions on AI trust and perceptions of existential threat, as well as delivering or collaborating on AI training sessions.

6 Conclusion

The AI market is continuing to grow and is increasingly being adopted for everyday use and critical applications due to the widespread benefits it can pose for both deployers and users. However, with this increased adoption has come increased awareness of the risks that AI can pose when it is left unchecked, as well as the potential benefits to society that the transformational potential of complex systems can have. Accordingly, the growing use of AI can be viewed through both a positive lens and negative lens. Indeed, applications of AI to areas such as healthcare, education, and scientific discovery are often seen as valuable to society, while the risk of bias, job displacement, alignment issues, and overestimation of the abilities of AI can make AI seem detrimental to society, sometimes leading to AI being depicted as an existential risk to humanity. However, it is not as simple as AI being categorically positive or negative; some use cases involve trade-offs where positives compete with uncovered hazards. For instance, self-driving vehicles promise increased accessibility but also pose unforeseen safety risks if deployed hastily. Overall, Artificial Intelligence is best served by an even-handed debate that resists false binaries. As this technology continues maturing, maintaining pragmatic expectations will allow us to maximize its advantages while proactively mitigating its dangers.

This paper's exploration of some of the myths around AI's current and prospective capability is an attempt at calming anxiety where common fiction has aroused it. Through the synthesis of research, this paper aimed to shed light on some actions that can be taken by various stakeholders in order to dispel fears and mitigate risks while encouraging ethical adoption. Specifically, we suggest that overcoming existential fears of AI requires the development of regulations and ethical guidelines to address issues such as the lack of transparency and accountability, self-governance in the lack of codified requirements, and awareness from users to ensure that systems are being used in an appropriate way. We also call for researchers to continue to investigate the changing perceptions of AI as self-governance becomes the norm and AI-specific legislation snowballs. As such, an interdisciplinary approach is crucial in this process and should be used to inform research around developing best practices for responsible AI development and use. Through collaborating with technical, social, and policy experts on the right messaging around AI developments, including where and when to share those messages, we can work towards building a positive and informed public perception of AI.

Author contributions Conceptualisation (EK, AH, SL), writing (EK, AH, SL), editing (AH, SL).

Data availability This article did not rely on any datasets.

Declarations

Competing interests The authors have no conflicts of interests to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Sturm, T.P.: Will Robots Destroy Us? Teaching Students About Technological Implications (2001)
2. Future of Life Institute (2023) Pause Giant AI Experiments: An Open Letter - Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Accessed 8 Sep 2023
3. Turing, A.M.: Computing machinery and intelligence. Mind **59**, 433–460 (1950)
4. Haenlein, M., Kaplan, A.: A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. Calif. Manag. Rev. **61**, 5–14 (2019). https://doi.org/10.1177/0008125619864925
5. Hunter, B., Hindocha, S., Lee, R.W.: The role of artificial intelligence in early cancer diagnosis. Cancers **14**, 1524 (2022). https://doi.org/10.3390/CANCERS14061524
6. Jackulin, C., Murugavalli, S.: A comprehensive review on detection of plant disease using machine learning and deep learning approaches. Meas. Sens. **24**, 100441 (2022). https://doi.org/10.1016/J.MEASEN.2022.100441
7. Adam, M., Wessel, M., Benlian, A.: AI-based chatbots in customer service and their effects on user compliance. Electron. Mark. **31**, 427–445 (2021). https://doi.org/10.1007/S12525-020-00414-7/FIGURES/7
8. IBM (2022) IBM Global AI Adoption Index 2022
9. IDC.: Worldwide Spending on AI-Centric Systems Will Pass $300 Billion by 2026, According to IDC (2022). https://www.i

dc.com/getdoc.jsp?containerId=prUS49670322. Accessed 21 Jul 2023

10. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women (2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 8 Jun 2021
11. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How We Analyzed the COMPAS Recidivism Algorithm. In: ProPublica (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 6 Jan 2023
12. (2021) Pruden v Lemonade Inc
13. (2022) Huskey v State Farm Fire & Casualty Company
14. (2023) Mobley v. Workday, Inc.
15. Kazim, E., Koshiyama, A.S.: A high-level overview of AI ethics. Patterns **2**, 100314 (2021). https://doi.org/10.1016/J.PATTER.2021.100314
16. Kazim, E., Koshiyama, A.: AI assurance processes. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3685087
17. Koshiyama, A., Kazim, E., Treleaven, P., et al.: Towards Algorithm Auditing A Survey on Managing Legal. SSRN Electronic Journal, Ethical and Technological Risks of AI, ML and Associated Algorithms (2021). https://doi.org/10.2139/SSRN.3778998
18. The New York City Council (2021) Int 1894–2020
19. European Commission.: Proposal for a regulation laying down harmonised rules on artificial intelligence (2021)
20. Araujo, T., Helberger, N., Kruikemeier, S., de Vreese, C.H.: In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc. **35**, 611–623 (2020). https://doi.org/10.1007/s00146-019-00931-w
21. Karger, E., Rosenberg, J., Jacobs, Z., et al.: Forecasting existential risks evidence from a long-run forecasting tournament (2023)
22. Cave, S., Coughlan, K., Dihal, K.: "Scary robots." In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. ACM, New York, NY, USA, pp 331–337 (2019)
23. Shoss, M.K., Ciarlante, K.: Are robots/AI viewed as more of a workforce threat in unequal societies? Evid. Eurobarometer Surv. (2022). https://doi.org/10.1037/tmb0000078.supp
24. OECD.: Recommendation of the Council on Artificial Intelligence (2019). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Accessed 7 Aug 2023
25. European Parliament, Council of the European Union (2024) Regulation (EU) 2024/1689
26. California Legislature.: AB-331 Automated decision tools (2023). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB331. Accessed 24 Jul 2024
27. National Institute of Standards and Technology.: AI Risk Management Framework (2023). https://www.nist.gov/itl/ai-risk-management-framework. Accessed 24 Jul 2024
28. Innovation Science and Economic Development Canada.: Artificial Intelligence and Data Act (2022). https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act. Accessed 24 Jul 2024
29. Council of Europe.: What's AI? (2018). https://www.coe.int/en/web/artificial-intelligence/what-is-ai. Accessed 23 Jul 2023
30. Connecticut General Assembly.: SB1103: An act concerning artificial intelligence, automated decision-making and personal data privacy (2023). https://www.cga.ct.gov/2023/act/Pa/pdf/2023PA-00016-R00SB-01103-PA.PDF. Accessed 24 Jul 2024
31. (2017) OECD Digital Economy Outlook 2017. OECD
32. Hanson Robotics.: Sophia - Hanson Robotics (2020). https://www.hansonrobotics.com/sophia/. Accessed 7 Aug 2023
33. European Parliament.: Civil Law Rules on Robotics. In: 2017 (2017). https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html. Accessed 7 Aug 2023
34. Scheel, P.D.: Robotics in industry: a safety and health perspective. Prof. Saf. Saf. **38**, 28 (1998)
35. Russell, S.: Human Compatible: AI and the Problem of Control. Allen Lane (2019)
36. Bostrom, N.: Superintelligence: Paths, Dangers. Oxford University Press, Strategies (2014)
37. Gabriel, I.: Artificial intelligence, values, and alignment. Minds Mach (Dordr) **30**, 411–437 (2020). https://doi.org/10.1007/S11023-020-09539-2/METRICS
38. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on Facebook. Science **348**, 1130–1132 (1979). https://doi.org/10.1126/science.aaa1160
39. Barocas, S., Selbst, A.: Big data's disparate impact. Calif. Law Rev. **104**, 671–732 (2016). https://doi.org/10.15779/Z38BG31
40. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of Machine Learning Research (2018), pp. 1–15
41. Ntoutsi, E., Fafalios, P., Gadiraju, U., et al.: Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **10**, e1356 (2020). https://doi.org/10.1002/WIDM.1356
42. Bigman, Y.E., Wilson, D., Arnestad, M.N., et al.: Algorithmic discrimination causes less moral outrage than human discrimination. J. Exp. Psychol. Gen. (2022). https://doi.org/10.1037/xge0001250
43. Capraro, V., Lentsch, A., Acemoglu, D., et al.: The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus (2024). https://doi.org/10.1093/PNASNEXUS/PGAE191
44. Wachter, S., Mittelstadt, B., Russell, C.: Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Comput. Law Secur. Rev.. Law Secur. Rev. **41**, 105567 (2021). https://doi.org/10.1016/J.CLSR.2021.105567
45. Mitchell, S., Potash, E., Barocas, S., et al.: Algorithmic fairness: Choices, assumptions, and definitions. Annu. Rev. Stat. Appl. **8**, 141–163 (2021). https://doi.org/10.1146/annurev-statistics-042720-125902
46. Ryan, A.M., Ployhart, R.E.: A Century of selection. Annu. Rev. Psychol.. Rev. Psychol. **65**, 693–717 (2014). https://doi.org/10.1146/annurev-psych-010213-115134
47. Schmidt. F.L., Oh, I.-S., Shaffer, J.A.: The validity and utility of selection methods in personnel psychology: practical and theoretical Implications of 100 Years (2016)
48. Schmidt, F.L., Hunter, J.E.: The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. Psychol. Bull. **124**, 262–274 (1998). https://doi.org/10.1037/0033-2909.124.2.262
49. Society for Industrial and Organizational Psychology.: Principles for the validation and use of personnel selection procedures, 5th ed (2018)
50. LaGrandeur, K.: The consequences of AI hype. AI Ethics **1**, 1–4 (2023). https://doi.org/10.1007/S43681-023-00352-Y
51. Barclay, P., Willer, R.: Partner choice creates competitive altruism in humans. Proc. Biol. Sci. **274**, 749–753 (2007). https://doi.org/10.1098/rspb.2006.0209
52. Fenn, J., Raskino. J.: Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time. Harvard Business Review Press (2008)
53. Cave, S., Craig, C., Dihal, K., et al.: Portrayals and perceptions of AI and why they matter. R. Soc. (2018). https://doi.org/10.17863/CAM.34502
54. Kruger, J., Dunning, D.: Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J. Pers. Soc. Psychol. **77**, 1121–1134 (1999). https://doi.org/10.1037/0022-3514.77.6.1121
55. Pennycook, G., Rand, D.G.: Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than

by motivated reasoning. Cognition **188**, 39–50 (2019). https://doi.org/10.1016/j.cognition.2018.06.011

56. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behav. Brain Sci.. Brain Sci. **40**, e253 (2017). https://doi.org/10.1017/S0140525X16001837
57. McKinsey & Company.: What the future of work will mean for jobs, skills, and wages: Jobs lost, jobs gained| McKinsey (2017). https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages. Accessed 21 Jul 2023
58. PricewaterhouseCoopers.: Will robots really steal our jobs? (2018)
59. Xu, M., David, J.M., Kim, S.H.: The fourth industrial revolution: opportunities and challenges. Int. J. Financial Res. (2018). https://doi.org/10.5430/ijfr.v9n2p90
60. The World Bank.: Unemployment, total (% of total labor force) (modeled ILO estimate)| Data (2022). https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?end=2022&start=2019&view=chart. Accessed 24 Oct 2023
61. World Economic Forum.: The Future of Jobs Report 2023
62. Reid, H.: IKEA bets on remote interior design as AI changes sales strategy. In: Reuters (2023). https://www.reuters.com/technology/ikea-bets-remote-interior-design-ai-changes-sales-strategy-2023-06-13/. Accessed 21 Jul 2023
63. Noy, S., Zhang, W.: Experimental evidence on the productivity effects of generative artificial intelligence. Science **381**, 187–192 (1979). https://doi.org/10.1126/SCIENCE.ADH2586
64. Gennatas, E.D., Chen, J.H.: Artificial intelligence in medicine: past, present, and future. In: Artificial Intelligence in Medicine. Elsevier, pp 3–18 (2021)
65. Jiang, F., Jiang, Y., Zhi, H., et al.: Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. **2**, 230–243 (2017). https://doi.org/10.1136/svn-2017-000101
66. Yip, M., Das, N.: Robot autonomy for surgery, pp. 281–313 (2018)
67. Haleem, A., Javaid, M., Singh, R.P., Suman, R.: Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens. Int. **2**, 100117 (2021). https://doi.org/10.1016/j.sintl.2021.100117
68. Choudhury, A., Asan, O.: Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med. Inform. **8**, e18599 (2020). https://doi.org/10.2196/18599
69. Palmisciano, P., Jamjoom, A.A.B., Taylor, D., et al.: Attitudes of patients and their relatives toward artificial intelligence in neurosurgery. World Neurosurg. **138**, e627–e633 (2020). https://doi.org/10.1016/j.wneu.2020.03.029
70. Márquez-Neila, P., Fisher, C., Sznitman, R., Heng, K.: Supervised machine learning for analysing spectra of exoplanetary atmospheres. Nat. Astron. **2**, 719–724 (2018). https://doi.org/10.1038/s41550-018-0504-2
71. Wang, H., Fu, T., Du, Y., et al.: Scientific discovery in the age of artificial intelligence. Nature **620**, 47–60 (2023). https://doi.org/10.1038/s41586-023-06221-2
72. Duan, Y., Edwards, J.S., Dwivedi, Y.K.: Artificial intelligence for decision making in the era of big data– evolution, challenges and research agenda. Int. J. Inf. Manag.Manag. **48**, 63–71 (2019). https://doi.org/10.1016/J.IJINFOMGT.2019.01.021
73. Rahal, C., Verhagen, M., Kirk, D.: The rise of machine learning in the academic social sciences. AI Soc. **39**, 799–801 (2024). https://doi.org/10.1007/s00146-022-01540-w
74. De Bie, T., De Raedt, L., Hernández-Orallo, J., et al.: Automating data science. Commun. ACM. ACM **65**, 76–87 (2022). https://doi.org/10.1145/3495256
75. Varadi, M., Anyango, S., Deshpande, M., et al.: AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. **50**, D439–D444 (2022). https://doi.org/10.1093/NAR/GKAB1061
76. Yang, W., Wei, Y., Wei, H., et al.: Survey on explainable AI: from approaches, limitations and applications aspects. Hum. Centric Intell. Syst. **3**, 161–188 (2023). https://doi.org/10.1007/S44230-023-00038-Y
77. Tuia, D., Kellenberger, B., Beery, S., et al.: Perspectives in machine learning for wildlife conservation. Nat. Commun.Commun. **13**, 792 (2022). https://doi.org/10.1038/s41467-022-27980-y
78. Joshi, M.: Adaptive learning through artificial intelligence. SSRN Electron. J. (2023). https://doi.org/10.2139/ssrn.4514887
79. Clark, D.: Artificial intelligence for learning: how to use AI to support employee development. Kogan Page Publishers (2020)
80. Goel, A.: AI-Powered Learning: Making Education Accessible, Affordable, and Achievable (2020)
81. Alessandro, G., Dimitri, O., Cristina, B., Anna, M.: The emotional impact of generative AI: negative emotions and perception of threat. Behav. Inf. Technol. (2024). https://doi.org/10.1080/0144929X.2024.2333933
82. Ipsos.: Ipsos finds doctors remain wary over patient use of health data, but are excited about AI in diagnosis| Ipsos (2023). https://www.ipsos.com/en-uk/ipsos-finds-doctors-remain-wary-over-patient-use-health-data-are-excited-about-ai-diagnosis. Accessed 22 Sep 2023
83. Tversky, A., Kahneman, D.: Judgment under uncertainty: heuristics and biases. Science **185**, 1124–1131 (1979). https://doi.org/10.1126/science.185.4157.1124
84. Kahneman, D.: Thinking, Fast and Slow. Penguin (2011)
85. Ellsberg, D.: Risk, ambiguity, and the savage axioms. Q. J. Econ. **75**, 643 (1961). https://doi.org/10.2307/1884324
86. Malle, B.F., Guglielmo, S., Monroe, A.E.: A theory of blame. Psychol. Inq. **25**, 147–186 (2014). https://doi.org/10.1080/1047840X.2014.877340
87. Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. Science **211**, 453–458 (1979). https://doi.org/10.1126/science.7455683
88. Tversky, A., Kahneman, D.: Availability: a heuristic for judging frequency and probability. Cogn. Psychol.. Psychol. **5**, 207–232 (1973). https://doi.org/10.1016/0010-0285(73)90033-9
89. Brauner, P., Hick, A., Philipsen, R., Ziefle, M.: What does the public think about artificial intelligence? A criticality map to understand bias in the public perception of AI. Front. Comput. Sci (2023). https://doi.org/10.3389/FCOMP.2023.1113903
90. European Parliament.: Artificial intelligence: How does it work, why does it matter, and what can we do about it? (2020). https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641547. Accessed 22 Jul 2024
91. Vaccari, C., Chadwick, A.: Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Soc. Media Soc. (2020). https://doi.org/10.1177/2056305120903408/ASSET/IMAGES/LARGE/10.1177_2056305120903408-FIG2.JPEG
92. Edelman, A.: States turn their attention to regulating AI and deepfakes as 2024 kicks off (2024). https://www.nbcnews.com/politics/states-turn-attention-regulating-ai-deepfakes-2024-rcna135122. Accessed 7 Feb 2024
93. Whitney, L.: ChatGPT is no longer as clueless about recent events (2023). https://www.zdnet.com/article/chatgpt-is-no-longer-as-clueless-about-recent-events/. Accessed 7 Feb 2024
94. Humphreys, D., Koay, A., Desmond, D., Mealy, E.: AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business. AI Ethics **2024**, 1–14 (2024). https://doi.org/10.1007/S43681-024-00443-4
95. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., et al.: Explanations can reduce overreliance on AI systems during

decision-making. Proc. ACM Hum. Comput. Interact. **7**, 1–38 (2023). https://doi.org/10.1145/3579605

96. Beauchene, V., de Bellefonds, N., Duranton, S., Mills, S.: AI at work: what people are saying (2023). https://www.bcg.com/publications/2023/what-people-are-saying-about-ai-at-work. Accessed 22 Sep 2023
97. Ka, C., Chan, Y., Hu, W.: Students' voices on generative AI: perceptions. Benefits Challenges High. Educ. (2023). https://doi.org/10.1186/s41239-023-00411-8
98. Horodyski, P.: Recruiter's perception of artificial intelligence (AI)-based tools in recruitment. Comput. Hum. Behav. Rep. **10**, 100298 (2023). https://doi.org/10.1016/J.CHBR.2023.100298
99. Gillespie, N., Lockey, S., Curtis, C., et al.: Trust in artificial intelligence: a global study (2023)
100. Cazorla, M., González-Calatayud, V., Almaraz-López, C., et al.: Comparative study of the attitudes and perceptions of university students in business administration and management and in education toward artificial intelligence. Educ. Sci. **13**, 609 (2023). https://doi.org/10.3390/EDUCSCI13060609
101. Said, N., Potinteu, A.E., Brich, I., et al.: An artificial intelligence perspective: How knowledge and confidence shape risk and benefit perception. Comput. Hum. Behav.. Hum. Behav. **149**, 107855 (2023). https://doi.org/10.1016/J.CHB.2023.107855
102. Nader, K., Toprac, P., Scott, S., Baker, S.: Public understanding of artificial intelligence through entertainment media. AI Soc. **1**, 1–14 (2022). https://doi.org/10.1007/S00146-022-01427-W/FIGURES/18
103. Nguyen, D., Hekman, E.: The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation. AI Soc. **1**, 1–15 (2022). https://doi.org/10.1007/S00146-022-01511-1/FIGURES/10
104. Datta, P., Whitmore, M., Nwankpa, J.K.: A perfect storm: social media news. Psychol. Biases AI (2021). https://doi.org/10.1145/3428157
105. von Eschenbach, W.J.: Transparency and the black box problem: why we do not trust AI. Philos. Technol. **34**, 1607–1622 (2021). https://doi.org/10.1007/s13347-021-00477-0
106. Choung, H., David, P., Ross, A.: Trust in AI and its role in the acceptance of AI technologies. Int. J. Hum. Comput. Interact. Comput. Interact. **39**, 1727–1739 (2023). https://doi.org/10.1080/10447318.2022.2050543
107. Dupont. J., Baron, D., Price, A., et al.: What does the public think about AI? (2024)
108. Yu, L., Li, Y.: Artificial intelligence decision-making transparency and employees' trust: the parallel multiple mediating effect of effectiveness and discomfort. Behav. Sci.. Sci. **12**, 127 (2022). https://doi.org/10.3390/bs12050127
109. Langer, M., Hunsicker, T., Feldkamp, T., et al.: "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems? In: CHI Conference on Human Factors in Computing Systems, pp. 1–28. ACM, New York, NY, USA (2022)
110. Langer, M., König, C.J., Fitili, A.: Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. Comput. Hum. Behav.. Hum. Behav. **81**, 19–30 (2018). https://doi.org/10.1016/j.chb.2017.11.036
111. X.: Twitter's Recommendation Algorithm (2023). https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm. Accessed 20 Jul 2024
112. HireVue.: Explainability Statement (2022)
113. Roski, J., Maier, E.J., Vigilante, K., et al.: Enhancing trust in AI through industry self-governance. J. Am. Med. Inform. Assoc. **28**, 1582–1590 (2021). https://doi.org/10.1093/jamia/ocab065
114. Winfield, A.F.T., Jirotka, M.: Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **376**, 20180085 (2018). https://doi.org/10.1098/rsta.2018.0085
115. Bedué, P., Fritzsche, A.: Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. J. Enterp. Inf. Manag.Enterp. Inf. Manag. **35**, 530–549 (2022). https://doi.org/10.1108/JEIM-06-2020-0233
116. Koshiyama, A., Kazim, E., Treleaven, P., et al.: Towards algorithm auditing: managing legal, ethical and technological risks of AI. R Soc Open Sci, ML and associated algorithms (2024). https://doi.org/10.1098/rsos.230859
117. Tamò-Larrieux, A., Guitton, C., Mayer, S., Lutz, C.: Regulating for trust: Can law establish trust in artificial intelligence? Regul. Gov. **18**, 780–801 (2024). https://doi.org/10.1111/rego.12568
118. European Commission.: AI Pact (2024). https://digital-strategy.ec.europa.eu/en/policies/ai-pact. Accessed 20 Jul 2024
119. UNESCO.: UNESCO and European Judicial Training Network partner to train judges on Artificial Intelligence and Rule of Law (2024)
120. Lemay, D.J., Basnet, R.B., Doleck, T.: Examining the relationship between threat and coping appraisal in phishing detection among college students. J. Internet Serv. Inf. Secur. **10**, 39–49 (2020). https://doi.org/10.22667/JISIS.2020.02.29.038
121. Office for Students.: New analysis shows over 7,600 students have enrolled on AI and data science courses to tackle digital skills gaps (2024). https://www.officeforstudents.org.uk/news-blog-and-events/press-and-media/new-analysis-shows-over-7-600-students-have-enrolled-on-ai-and-data-science-courses-to-tackle-digital-skills-gaps/. Accessed 20 Jul 2024

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