The impact of generative artificial intelligence on socioeconomic inequalities and policy making

Valerio Capraro[1,*], Austin Lentsch[2], Daron Acemoglu[3], Selin Akgun[4], Aisel Akhmedova[4], Ennio Bilancini[5], Jean-François Bonnefon[6], Pablo Brañas-Garza[7], Luigi Butera[8], Karen M. Douglas[9], Jim A.C. Everett[10], Gerd Gigerenzer[11], Christine Greenhow[12], Daniel A. Hashimoto[13,14], Julianne Holt-Lunstad[15], Jolanda Jetten[16], Simon Johnson[17], Chiara Longoni[18], Pete Lunn[19], Simone Natale[20], Iyad Rahwan[21], Neil Selwyn[22], Vivek Singh[13], Siddharth Suri[23], Jennifer Sutcliffe[4], Joe Tomlinson[24], Sander van der Linden[25], Paul A. M. Van Lange[26], Friederike Wall[27], Jay J. Van Bavel[28, 29], Riccardo Viale[30]

[1] Department of Psychology, University of Milan-Bicocca, Italy. [2] Department of Economics, MIT, USA. [3] Institute Professor and Department of Economics, MIT, USA. [4] College of Education, Michigan State University. [5] IMT School of Advanced Studies Lucca, Italy. [6] Toulouse School of Economics, France. [7] Department of Economics, Loyola Andalucia University, Spain. [8] Copenhagen Business School, Denmark. [9] School of Psychology, University of Kent, UK. [10] School of Psychology, University of Kent, UK. [11] Harding Center for Risk Literacy, University of Potsdam, Germany. [12] College of Education, Michigan State University. [13] Penn Computer Assisted Surgery and Outcomes Laboratory, Department of Surgery, Perelman School of Medicine, University of Pennsylvania. [14] Department of Computer and Information Science, School of Engineering and Applied Science, University of Pennsylvania. [15] Department of Psychology and Neuroscience, Brigham Young University, USA. [16] School of Psychology, University of Queensland, Australia. [17] MIT Sloan School of Management, USA. [18] Department of Marketing, Bocconi University, Italy. [19] Economic & Social Research Institute, Dublin, Ireland. [20] Department of Humanities, University of Turin, Italy. [21] Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany. [22] Faculty of Education, Monash University, Australia. [23] Microsoft, USA. [24] York Law School, United Kingdom. [25] Department of Psychology, University of Cambridge, UK. [26] Department of Experimental and Applied Psychology, Vrije Universiteit, Amsterdam, The Netherlands. [27] Department of Management Control and Strategic Management, University of Klagenfurt, Austria. [28] Department of Psychology & Center for Neural Science, New York University, USA. [29] Norwegian School of Economics, Bergen, Norway. [30] Department of Economics, University of Milan-Bicocca, Italy.

* Corresponding author: valerio.capraro@unimib.it

Abstract

Generative artificial intelligence, including chatbots like ChatGPT, has the potential to both exacerbate and ameliorate existing socioeconomic inequalities. In this article, we provide a state-of-the-art interdisciplinary overview of the probable impacts of generative AI on four critical domains: work, education, health, and information. Our goal is to warn about how generative AI could worsen existing inequalities while illuminating directions for using AI to resolve pervasive social problems. Generative AI in the workplace can boost productivity and create new jobs, but the benefits will likely be distributed unevenly. In education, it offers personalized learning but may widen the digital divide. In healthcare, it improves diagnostics and accessibility but could deepen pre-existing inequalities. For information, it democratizes content creation and access but also dramatically expands the production and proliferation of misinformation. Each section covers a specific topic, evaluates existing research, identifies critical gaps, and recommends research directions. We conclude with a section highlighting the role of policymaking to maximize generative AI's potential to reduce inequalities while mitigating its harmful effects. We discuss strengths and weaknesses of existing policy frameworks in the European Union, the United States, and the United Kingdom, observing that each fails to fully confront the socioeconomic challenges we have identified. We contend that these policies should promote shared prosperity through the advancement of generative AI. We suggest several concrete policies to encourage further research and debate. This article emphasizes the need for interdisciplinary collaborations to understand and address the complex challenges of generative AI.

Introduction

'The rise of powerful AI will be either the best, or the worst thing, ever to happen to humanity. We do not yet know which." - Stephen Hawking, 2016

Recent and future advances in generative Artificial Intelligence (AI) represent a shift in the capability of AI systems to solve problems previously thought unsolvable (Bucker et al., 2023). Countless techno-optimists predict a utopian future where machines can perform an ever-increasing number of tasks—but humans remain in control, the gains from prosperity are shared throughout society, and we all enjoy lives with less work and more leisure. On the other hand, less optimistic forecasts suggest that we are headed toward a dystopian future where machines not only replace humans in the workplace, but also surpass human capability and oversight, destabilize institutions and destroy livelihoods—and perhaps even cause the downfall of humanity (Campbell, 2023; Andreessen, 2023; Bostrom, 2003).

Melvin Kranzberg, a prominent scholar in the history of technology, in a presidential address to his field, defined "Kranzberg's Laws", the first of which states that *"Technology is neither good nor bad; nor is it neutral"* (Kranzberg, 1985). This principle suggests that technologies like generative AI could have either negative or positive impacts (or both) on society, though they are not inherently predestined toward either. This article aims to outline some of these effects, with the hope of guiding society to harness AI's positive effects while avoiding the negative ones. Generative AI will impact virtually every facet of society. In this article, we speculate on the impact of AI on socioeconomic inequalities in four key areas: work, education, health, and information. For each domain, we explore current research and suggest broad directions for future exploration.

In the workplace, generative AI could increase productivity and promote shared prosperity, especially when used to complement human efforts and create new well-paid jobs—offsetting workplace automation with new, value-adding task creation. However, the benefits and costs will likely be distributed unevenly across firm sizes, sectors, and worker demographics. In education, generative AI promises personalized learning experiences, potentially bridging educational gaps. However, it also raises concerns about the digital divide and equal access to these advanced tools. The health sector could greatly benefit from AI's diagnostic and predictive capabilities, improving patient outcomes and making healthcare more accessible. Yet, there is the risk of deepening existing inequalities of care and access, especially for under-resourced and marginalized communities. The information domain, too, is set to be radically transformed. Generative AI can democratize content creation and access but also leads to challenges such as increased misinformation and erosion of trust in digital content.

We conclude with an examination of the role of policymaking in the age of artificial intelligence. We discuss the pros and cons of the current policy approaches in the European Union, the United States, and the United Kingdom, noting that all fall short in adequately addressing the socioeconomic risks that we identify. We argue that policies must be designed to mitigate the potential problems posed by AI, aiming for an equitable distribution of benefits across society. We propose several explicit policy recommendations that could be discussed in public debate and research endeavors. This includes strategies to prevent job market inequalities, initiatives to bridge the digital divide in education and healthcare, and measures to combat AI-generated misinformation. The ultimate goal should be to harness the potential of generative AI in ways that favor human flourishing, striking a balance between technological advancement and societal well-being.

Impact on work

Previous waves of digital technologies have contributed to increased inequality. Some of these technologies, like personal computers, have been complementary mostly to more-educated workers (Autor et al., 1998; Goldin and Katz, 2008; Autor et al., 2003; Autor, 2019), while others, like industrial robots, have been used to automate repetitive or systematic job tasks that are often performed by less-educated workers (Acemoglu and Restrepo, 2022a, 2022b; Autor et al., 2003, Restrepo 2023). Together, the upside for more-educated workers and downside for less-educated workers have magnified the distributional consequences of technological innovation.

The current trend in AI emphasizes automation. While some amount of this is unavoidable, the displacement of labor by “so-so technologies” (e.g., self-checkout kiosks or automated phone systems) that offer little or no productivity gain, along with diminished worker voice due to intensified monitoring and surveillance, can be harmful to long-run productivity and other social goals like job satisfaction (Acemoglu and Restrepo, 2022a, 2019). Although new technologies can boost productivity in some areas (e.g., Brynjolfsson and McAfee, 2016), the productivity gains from those technologies have often fallen well below expectations, especially when the focus has been on replacing work instead of augmenting pre-existing worker capabilities or developing new ones (e.g., Acemoglu et al., 2016; Acemoglu and Johnson, 2023).

Many businesses and researchers tend to focus on automating work instead of creating new job tasks and enabling workers to build new skills. Reasons for this may include hopes for cost-savings, eliminating demanding workers, reducing uncertainty, increasing control, and (to some extent) following the dominant intellectual paradigm of Silicon Valley that focuses on developing AI agents to mimic or surpass all human capabilities as quickly as possible (Acemoglu and Johnson, 2023; Acemoglu, Autor et al., 2023).

New technologies like AI should be oriented not so much toward replacing human problem-solving abilities, but rather toward enhancing them in a symbiotic relationship where machines are designed to complement human capabilities and humans can compensate for the weaknesses of machines (Licklider, 1960; Engelbart, 1995). This “pro-worker” or “human-complementary” path could contribute more to productivity growth and could help reduce economic inequality. The critical question we face in the new era of generative AI is whether this technology will primarily accelerate the existing trend of automation without the offsetting force of good-job creation—particularly for non-college educated workers—or whether it will instead introduce new value-adding tasks and well-paying jobs for workers with diverse skill sets and educational backgrounds.

There is cause for optimism: AI can complement workers by making them more efficient, helping them to produce higher quality work, or enabling them to take on new value-adding tasks (Acemoglu and Restrepo, 2018; Acemoglu, Ahmed et al., 2023; Korinek, 2023). Recent evidence indicates that this might be the case, especially for generative AI. Brynjolfsson and et al. (2023), for instance, consider the staggered implementation of a chat assistant by a Fortune 500 software company that provides business process software. The chat assistant monitored customer service chats and proposed real-time response suggestions to customer service agents. Agents had the option to use or ignore these suggestions. Access to the AI assistant increased productivity. Less-skilled or inexperienced workers were enabled to resolve around 34% more issues per hour, with average improvement across all workers measuring about 14% (the tool was less impactful for experienced and highly skilled workers). Agents using the tool with only two months of tenure performed as well as those without the tool who had more than six months of tenure.

Another study examined the impact of GPT-4 access on consultants' abilities to perform complex knowledge-intensive tasks. AI users were generally more productive and produced higher quality work. However, for tasks beyond the capabilities of GPT-4—specifically, tasks that involve imperfect information or omitted data, which require cross-referencing resources and leveraging experience-gained intuition to complete successfully—AI usage resulted in fewer correct solutions. Consultants with below-average performance improved by 43% with AI, while those above average improved by 17% (Dell'Acqua et al., 2023). This suggests that AI might reduce inequalities in performance among knowledge workers

Similar patterns have been observed in other studies. For instance, Peng and co-authors (2023) conducted a controlled experiment with GitHub Copilot, an AI-based programming assistant. Professional programmers were tasked to implement an HTTP server in JavaScript. Programmers with access to the AI copilot completed the task in 71 minutes on average, less than half the time of the control group's 161 minutes. The AI assistant provided the biggest boost to less-experienced and older programmers, as well as those coding more hours daily. In a controlled online experiment, people with access to ChatGPT completed a writing task faster and produced higher quality work (Noy and Zhang, 2023). As with the other studies, this reduced worker inequality by benefiting lower-ability workers more; moreover, it led to higher job satisfaction and self-efficacy.

These studies underscore the potential of generative AI to disproportionately boost productivity for workers with less experience or skill. This fundamentally differs from other recent waves of new technology (e.g., Internet, computers), which have overwhelmingly aided *highly skilled workers* much more than less-skilled workers. From the mid-19th century through the 1970s, the worker-displacing effects of automation were generally offset by the creation of new tasks, which allowed low-skilled workers to obtain new jobs and higher wages as technology evolved (Acemoglu and Johnson, 2023). By the 1980s, however, new task creation lagged behind automation, particularly for workers without a college education, adversely impacting wages and employment opportunities throughout the developed world (Acemoglu and Restrepo, 2019; Autor et al., 2022; Acemoglu and Johnson, 2023). Thus, the seeming “inverse skill-bias” of worker-complementary generative AI, benefitting *less-skilled workers* much more than highly skilled workers, could radically change how technology affects labor markets. This paradigm shift—toward a better balance of automation and augmentation—could perhaps counteract or even reverse the trends toward greater inequality observed over the last several decades.

Generative AI could also reduce inequality by reducing barriers to entry in the digital economy. For example, its translation capabilities can help overcome language barriers. This increased accessibility, in conjunction with trends toward diminishing geographic barriers, could have a compounding positive effect. Geographic barriers are challenging to remove — collaborating across many time zones is difficult and many firms prefer to adopt hybrid work, where employees come into the office several times per week, instead of fully virtual arrangements (Aksoy et al., 2023). Nonetheless, there has been a surge of interest in remote-enabled digital economy jobs, evidenced by online job searches. Notably, a substantial part of this increase originated from rural areas (Counts et al., 2022).

One of the more notable strengths of generative AI is its ability to parse and aggregate enormous amounts of information. This capability can equalize access to information and lower research costs by simplifying online search tasks. If a user wants to accomplish a complex task with a traditional search engine, they have to break that task up into pieces, issue search queries for each piece, read the web pages returned by the search engine, assess the representativeness of their gathered information, and then aggregate the results to solve the original problem. Generative search engines, on the other hand, can aggregate this information and return it to the user, requiring less bandwidth and fewer trips between the user and the system which would be helpful in lower resource environments. In addition to the time and cost savings, these tools could compensate for expertise by identifying trustworthy resources and extracting the consensus on any topic by simultaneously considering more information than human operators can retain. This approach could help users and small businesses in low-resource settings access information that has traditionally been available only in high-resource environments.

There are also ways in which recent advances in AI might exacerbate inequalities in the workplace. One concern is differential access to these new tools. The most widely available and accessible generative AI platforms still require additional technical inputs (e.g., internet access and internet-enabled devices) as well as training to optimize performance. Industries, firms, and workers that have not yet integrated the prerequisite technologies will struggle to take advantage of the expanded capabilities and consequent productivity and earnings upsides, likely falling (further) behind well-resourced competitors or coworkers.

The role of firm behavior and social context matters, as there could be backfire effects even from well-designed tools. For example, while the evidence discussed above suggests that the introduction of generative AI tools gives more of a boost to less-skilled workers than highly skilled workers, this equalizing force could be a way for workers to increase their earnings potential, if compensation is tied to capability. Instead, if firms exploit the higher interchangeability between workers ("why hire an expert copywriter if a less-skilled writer with an AI chatbot can do the same level of work?") these wage gains may never be realized. Similarly, it could be possible that a single expert is enabled to do the work of multiple experts or direct reports very quickly with generative tools—the enormous volume of search results for "how to build a website with ChatGPT in one minute" suggests that this is an anticipated use case. This could slash the talent requirements for many business endeavors, including producing coded deliverables, marketing copy, graphic design, data analysis, etc.

AI will likely have outsized impacts on U.S. workers with Bachelors' or Associates' degrees, compared to higher or lower levels of education (Septiandri et al., 2023). This effect could compound over time: if generative AI tools commodify expertise and reduce the returns to specialized skills, workers may no longer spend the time or resources to acquire greater levels of expertise, leading to lower levels of worker skill and overreliance on outsourcing to generative tools. These effects could cause greater competition at the (now larger) lower end of the skill distribution, further depressing wages. There could be further downsides to productivity if non-automatable job tasks would benefit from workers having acquired the sort of foundational knowledge that is now disincentivized.

Governments may play an important role in mitigating the risk of increased inequality and maximizing the productivity potential of new generative AI tools. Explicit policy suggestions are postponed to the "Policymaking in the age of artificial intelligence" section (see also Table 5). Table 1 reports a succinct summary of the main research directions on the impacts of generative AI in the workplace.

Future research directions
Investigate how AI can be designed and implemented to augment human skills and increase productivity, rather than to simply replace workers and forego the long-run productivity upsides of maximizing workers' contributions to production. Examine how AI can facilitate more access to economic opportunities, particularly through reducing language-related barriers and promoting remote work technologies that can democratize access to the digital economy. Conduct long-term studies to monitor the evolving impact of AI on the workforce, capturing both the immediate and delayed effects on work across educational and occupational strata. Explore how AI can be utilized in educational and training programs to encourage basic competency with generative AI tools and better-equip workers in vulnerable job sectors in anticipation of labor market changes. Research labor laws, taxation policies, and social support systems that could support workers displaced or disadvantaged by AI.

Table 1. Summary of the main research directions on the impact of generative AI in workplace environments.

Impact on education

Various forms of generative AI are beginning to enter education, from chatbots that guide students' learning to text and image generation tools for producing lesson content. The integration of generative AI into schools, colleges and universities offers various benefits, including the potential of skill-adaptive and personalized teaching, instantaneous feedback, and on-demand student guidance and support. These uses could be particularly effective in large class settings, with significant opportunity to scale-up implementation beyond the capabilities of traditional educational practices. Consequently, generative AI could bridge complex and persistent educational gaps.

A review of AI applications in education identified several use cases that produced higher test scores when students used personalized learning systems (Akgun & Greenhow, 2022). These systems, unlike traditional approaches like static worksheets with standardized questions, detect areas where students lack foundational understanding by adapting educational resources and tools to foster their development. Furthermore, assessment algorithms can expedite grading of written assessments, which supports students by offering timely feedback that can be applied immediately. These systems have the potential to improve learning outcomes among students with a broad spectrum of learning styles. Students themselves perceive AI as potentially beneficial to their education. College students reported that generative AI provided personalized learning, supported their writing and brainstorming, and assisted with research and analysis (Chan & Hu, 2023) . However, students also expressed concerns about the accuracy, privacy, and ethical implications of generative AI tools—including how this technology could adversely impact their personal development and career prospects.

Educational uses of generative AI pose a number of challenges. One is the perpetuation of biases and discrimination, potentially reinforcing racial or gender-based stereotypes during personalized learning, automated scoring, and admission processes (Akgun and Greenhow, 2022; Baker and Hawn, 2022; Bender et al., 2021; Morewedge et al., 2023). The data used to train AI models could suffer from bias, if those data are based on past human decision making (a notoriously biased process). An example is the translation bias observed in tools like Google Translate, where gender stereotypes are inadvertently perpetuated in language translations. Translating the phrase "she/he is a nurse" from Turkish (which is "genderless") to English (which is "gendered") yielded the feminine form (i.e., "she is a nurse"), while the phrase "she/he is a doctor" yielded the masculine form (i.e., "he is a doctor"; Johnson,2021). Failing to account for these biases could amplify inequalities and injustices, specifically towards historically marginalized groups.

Although human teachers may also be prone to bias and discrimination—and AI systems can theoretically be designed to be less biased than humans—simply introducing slightly less discriminatory technologies into classrooms is not a substitute for the goal of removing discrimination from school (Pasquale, 2020). Moreover, these systems should be designed with sufficient transparency for users to monitor for and identify potential biases to ensure that these tools effectively serve their intended purposes (Stoyanovich et al., 2020).

Generative AI may place increased burdens on teachers. In contrast to the idea that AI tools relieve teachers of repetitive and onerous work, there is growing concern that teachers have to engage in additional tasks “behind the scenes” (e.g., curating and filtering content, monitoring student-AI interactions, providing technical support) to ensure that AI tools are able to function in complex classroom settings (Selwyn et al., 2023). This could exacerbate a generational divide among educators, as younger teachers may be more adept with new technology than older teachers. Furthermore, there could be unintended consequences of generative AI on student learning—for example, if students become overly reliant on extensive support from AI tools, this could undermine the capacity of students to work or think independently. Questions also arise about the accuracy of AI-generated content and the new skills that students must acquire to work effectively with AI systems, such as the ability to evaluate AI-generated content.

The current debate about the role of generative AI, from primary schools to universities, revolves around whether generative AI should be banned, permitted under only some cases, or generally allowed as assistance for teachers and students. For instance, the New York City education department and Chinese universities have banned generative AI (Elsen-Rooney, 2023; Liu et al., 2023), while the Berlin universities recommended its use in certain scenarios. A growing literature recommends the use of generative AI for teacher and student assistance within the traditional curricula (e.g., Chiu, 2023).

We argue that these approaches are limited in vision. A more forward-thinking approach would involve a curricular revolution to redefine the skills and competencies necessary to effectively utilize generative AI. Calculators did not remove the need for students to learn the properties of algebra and develop mathematical reasoning. Similarly, the internet did not eliminate the need for careful research and fact-checking; in fact, it increased this need, as online information is frequently incorrect or incomplete (Lazer et al., 2018). In the same vein, generative AI will not eliminate the need to learn effective thought organization, writing, and critical thinking skills. Therefore, curricula must teach how to successfully describe and share ideas, both with and without assistance from generative AI. In addition, they need to emphasize the development of critical-thinking skills, fact-checking abilities, an understanding of how generative AI tools function, and appropriate rules of interaction—including by refraining from anthropomorphizing (and thus misunderstanding) these tools (Kasneci et al., 2023).

More specifically, the text-production abilities of generative AI present an opportunity to teach students critical thinking. This will enable them to evaluate the argument and structure of the generated text and also to write intelligent prompts for generative AI. This skill should be recognized and assessed by educators. The output of generative AI is much more variable than other educational technologies (e.g., calculators); therefore, developing these critical thinking abilities and prompt-engineering skills is fundamental. Another crucial skill is the ability to fact-check generative AI outputs. Fact-checking skills are not taught sufficiently in schools. For instance, among more than 3,000 U.S. high school students and undergraduates, 96 percent did not know how to evaluate the trustworthiness of websites (Breakstone et al., 2021). These fact-checking abilities include smart heuristics such as lateral reading; i.e., the practice of navigating away from an unfamiliar website to verify the reliability of its information by consulting other external sources (McGrew, 2024). A toolbox of similar fact-checking heuristics needs to be developed or remediated for AI-generated content. Lastly, understanding the nature of large language models, which are statistical machines that calculate correlations between words, is essential. Only in this way can students understand the potential and limits of generative AI, rather than assuming that contemporary generative AI can "think" or "comprehend" like humans.

The adaptation of curricula is challenging, but essential. Without such changes, teachers and students may use generative AI merely as an automated assistance tool. This would forego the opportunity to develop higher-order cognitive skills, such as critical judgment and fact-checking, that generative AI itself cannot reliably perform. The result would be a likely decline in higher-order cognitive skills, especially in segments of the population that will use these tools in a more mechanical, less analytical manner. The role of governments in integrating generative AI into the education sector is crucial. We will discuss potential policy recommendations in the final section (see Table 5). Table 2 summarizes the main research directions.

Future research directions
Investigate how curricula can be redesigned to include generative AI as a tool for enhancing learning while also teaching students to critically engage with and understand this technology. Study effective training methods for teachers to integrate AI tools into their teaching practices and identify the additional support required to manage these technologies in the classroom. Explore strategies to ensure equitable access to AI educational tools, particularly for students in underprivileged or remote areas. Evaluate the long-term impacts of generative AI on student learning, teacher workloads, and educational outcomes. Examine how generative AI can be effectively used for personalized learning.

Table 2. Summary of the main research directions on the impact of generative AI on education.

Impact on Healthcare

Recent advances in AI techniques can democratize healthcare by making efficacious medical care more accessible and affordable. This is often achieved via augmenting human capacities and reducing workload: AI can support clinicians with diagnosis, screening, prognosis, and triaging, alleviating the burden on health practitioners and giving them the "gift of time" (Topol, 2019). For instance, a review of workplace burnout among healthcare providers identified electronic health record systems as a cause of increased stress due to insufficient documentation time, a high volume of patient communications, and negative perceptions by providers (del Carmen et al., 2019). In response, generative AI models have been suggested to aid in the completion of electronic health record-related tasks, reducing healthcare professionals' administrative demands (Patel and Lam, 2023).

AI-systems could also assist healthcare providers by analyzing and interpreting multimodal clinical data (e.g., photos, radiology images, and surgical videos) to provide relevant information to clinicians (The Lancet Regional Health – Europe, 2023). In one study, endoscopists reviewed colonoscopy videos with and without AI assistance. The results demonstrated that their decisions were influenced by AI, particularly when its advice was correct. This Bayesian-like integration of human and AI judgment led to superior performance compared to either alone, highlighting effective human-AI collaboration dynamics in medical decision-making (Reverberi et al., 2022). As a cautionary tale, other preliminary evidence finds that diagnostic performance of some expert physicians may not be improved by AI—and in fact may cause incorrect diagnoses in situations that otherwise would have been correctly assessed (Agarwal et al., 2023).

AI systems can also aid in "medical visual question answering"—analyzing medical images (like X-rays or MRI scans) and providing answers to specific questions about these images, typically by leveraging advanced image recognition and AI algorithms (Ren and Zhou, 2020). The current GPT-4 model demonstrates reasonable diagnostic accuracy in simple cases and can answer questions on standardized medical exams, though it struggles with diagnostically complex prompts (Kanjee et al., 2023). The totality of evidence suggests that more research is needed to understand when human-AI interactions are beneficial or detrimental to clinical practice, as well as appropriate training to avoid over-reliance of human physicians on AI-generated diagnostic suggestions.

Generative AI could also enable patients to manage their health more proactively through applications that patients can access outside of clinical settings. ChatGPT, for instance, has reasonable accuracy in answering common myths about cancer (Johnson et al., 2023). People trusted ChatGPT's answers to low-risk medical questions, though trust reportedly varied for questions with greater medical complexity (Nov et al., 2023). Furthermore, ChatGPT's answers to medical questions posted on Reddit's r/AskDocs were rated as higher quality and more empathetic than those of physicians 79% of the time (Ayers et al., 2023).

Conversational agents based on generative AI can also provide greater access to medical advice and simplify medical jargon. This may have positive downstream effects on inequality. Being part of a stigmatized group affects people's engagement and utilization of healthcare services. For example, when contextual cues made racial stereotypes salient, Black women were more likely to feel anxious in a healthcare setting than their white counterparts (Abdou and Fingerhut, 2014). More generally, there is evidence of considerable mistrust between health professionals and members of stigmatized groups (Cuevas et al., 2016; López-Cevallos et al., 2014). As a result, members of stigmatized minority groups are less likely to listen to, or trust, doctors who they perceive as outgroup members (Dovidio et al., 2008). For example, Black patients tend to be less satisfied with consultations, less likely to book an appointment, and have lower rates of medical compliance when they have consultations with white rather than Black physicians (Williams, 2005).

There are reasons to believe that AI-led healthcare will be more immune to pre-existing systemic biases or discriminatory practices than human-led healthcare. For example, health professionals are biased in their treatment of higher-weight patients (Rathbone et al., 2020). Incorporating AI-based tools into treatment decision making may lead to less bias if it can ameliorate these stereotypes and prejudices. Furthermore, interactions between members of stigmatized groups and the healthcare system might be more positive when the system is AI-led because their stigmatized status is not made salient in the interaction. This suggests that members of stigmatized groups could become more likely to engage with AI-led healthcare because they worry less about group- or identity-based factors affecting their treatment options (Hommel et al., 2012). However, it is important to recognize that many societal biases are baked into training datasets—often composed of human clinicians' decisions—and such biases are difficult to overcome.

Benefits aside, patients, medical providers, and those managing healthcare systems may be hesitant to adopt AI due to several psychological barriers. In fact, the impact of AI on clinical practice has been limited despite the growing number of AI tools (Yin et al., 2021; Aristidou et al., 2022). One key factor is public trust in AI technologies in healthcare (Quinn et al., 2021). For instance, patients may resist adoption because of misperceptions about AI, such as the belief that AI cannot account for a person's uniqueness as well as a human doctor (Longoni et al., 2019), or because of difficulty in holding AI accountable for mistakes (Promberger and Baron, 2006).

Another factor implicated in adoption hesitancy is the contrast between AI's opaqueness and the illusory perception that human decision-making is more transparent than AI. In reality, decisions made by human physicians or AI are probably equally unobservable to a patient—but because patients feel that they can understand the decision-making as explained by human providers, they ultimately penalize and resist the clinical use of AI (Cadario et al., 2021). The most recent versions of AI tools may be less susceptible to concerns about AI's inscrutability, since the iterative nature of newer generative AI tools may allow patients to ask follow-up questions in a more familiar, conversational format. It is possible that the back-and-forth supported by modern generative AI tools will empower patients with greater information about AI-driven decision-making, at which point patients may be better-equipped to decide whether to trust (or not trust) AI-generated medical recommendations.

Other challenges to AI adoption include pushback from healthcare practitioners—who may feel more comfortable with traditional methods of patient care, or who fear being replaced by machines—and from those managing healthcare systems, who might be reluctant to initiate costly and systemic changes until the usefulness of AI-integration is fully proven.

Insurance markets will also be impacted. Insurers could use AI to refine their practices, capturing a larger share of the surplus. This could lead to welfare losses for consumers. Today, it is not possible to determine highly accurate, individualized probabilities for the future health conditions of a particular insurant—insurance as a field relies instead on population-level probabilities, with some refinement from explicit risk factors. However, if generative AI allows companies to more accurately estimate this probability—for example, by incorporating information from unobservable factors that are identifiable only through advanced machine learning algorithms run on text-based claims, electronic health records, or other data—they might charge higher premiums to those at greater risk without offering reductions to those at lower risk.

AI could also enable insurers to reach currently uninsured groups, reducing inefficiencies and achieving market completeness. A concrete example of this is the use of Responsible Artificial Intelligence in healthcare to predict and prevent insurance claim denials, which could lead to significant cost savings and improved patient wellbeing (Johnson et al., 2021). Moreover, the application of AI by insurance companies might allow for a more accurate prediction of loss probabilities, thus reducing one of the industry’s most inherent problems, namely asymmetric information (Eling et al., 2022).

Generative AI may come to fulfill social needs for some people, which could have downstream effects on health. There is robust evidence linking social connectedness or lack thereof to long-term health outcomes (Holt-Lunstad, 2021, 2022; Van Lange and Columbus, 2021), including increased risk for chronic illnesses such as cardiovascular disease and stroke (Cené et al., 2022; Valtorta et al., 2016), type 2 diabetes, and dementia (Penninkilampi et al., 2018), as well as mortality from all causes (Leigh-Hunt et al., 2017; Wang et al., 2023; Holt-Lunstad et al., 2010). Generative AI can be used as a conversational companion, potentially replacing some human interaction. Indeed, digital proxies for social connection may, with increasing sophistication, mimic features of social connection, which could in turn decrease motivation to develop authentic human relationships. While digitally mediated forms of socializing (e.g., social media) have been utilized for years, there is increasing concern about the implications of these platforms for mental, social, and physical health, as highlighted by the U.S. Surgeon General and various studies (Twenge et al., 2022; U.S. Surgeon General, 2023; Valkenburg et al., 2022). One risk is that these features may relieve some of the tensions of human connection, leading people to preferentially spend more time with AI than humans or even form pseudo-social attachments to AI systems.

AI-based chatbots are insufficient stand-ins for customary human interactions (which is likely true), then many negative consequences could result. Humans are social beings, so our biological systems can become dysregulated when social needs are unmet, leading to poorer health (Beckes and Sbarra, 2022). Therefore, it is essential that some key elements of customary human interactions be retained – for example, research finds that relative to emails and other text-based interactions, those involving human voice boost social connection (Kumar and Epley, 2021). At the same time, AI-based chatbots could be useful to add social experiences for some individuals (while not completely replacing human-to-human interaction), particularly for those facing difficulties developing relationships on their own (who need “Vitamin S,” from Social contact, *see* Van Lange and Columbus, 2021), but are likely to be a poor or even dangerous replacement for human interaction writ large.

In sum, generative AI presents significant opportunities to alleviate inequalities in physical and mental health, in addition to augmenting healthcare providers’ capabilities. Additionally, it is crucial to ensure that generative AI are only designed to supplement, rather than replace, human social interactions. Excessive dependence on AI for social engagement could lead to various adverse outcomes, including social isolation and deteriorating mental and physical health. Table 3 outlines key areas for future research. In the concluding section, we offer policy recommendations designed to effectively integrate AI systems into healthcare frameworks, aiming to diminish healthcare disparities.

Future research directions
Research how AI can assist healthcare professionals in diagnosis, treatment planning, and patient monitoring. Investigate the use of AI to reduce the administrative burden on healthcare providers through efficient electronic health records management. Study how AI can contribute to the development of personalized medicine, adapting treatments to individual patient needs and reducing healthcare disparities. Investigate strategies to increase public trust and understanding of AI in healthcare. Research how AI can improve healthcare accessibility in underserved regions and populations, in both rural and urban areas. Investigate the potential of AI to facilitate social connections, particularly for individuals with difficulties in forming relationships, while also studying the potential risks of over-reliance on AI for social interaction.

Table 3. Summary of the main research directions on the impact of generative AI on healthcare.

Impact on (mis)information

Generative AI offers a broad spectrum of benefits in the information domain. One key advantage is the ability for personalization, where AI can tailor content to individual preferences, enhancing and customizing user experiences in areas such as education, entertainment, and news media. Language translation and localization capabilities of AI extend the reach of content globally, breaking language barriers and adapting material to different cultural contexts. Importantly, AI aids in making information more accessible, particularly for individuals with disabilities, by creating text-alternative formats like audio or simplified summaries. AI is also being explored to upscale and automate the fact-checking process, aiding the spread of accurate information (Hoes et al., 2023).

Concerningly, new generative AI technology and sophisticated machine learning techniques may also enable companies and platforms to collect and deploy excessive amounts of information about individuals. This will enable exploitation of consumers' biases or vulnerabilities in order to capture more of the consumer surplus via price discrimination or violations of consumer privacy in processing and using that data without proper consent, leading to what has been named "the age of surveillance capitalism" (Acemoglu, 2024; Zuboff, 2023). A dominant firm organization model has emerged from these data monopolies, where internet platforms earn income by aiming to optimally market digital advertisements to users (Acemoglu and Johnson, 2023). This sort of business strategy necessarily places a premium on user attention, which has led companies to deploy AI and machine learning techniques in ways intended to prolong user engagement, often to the detriment of users' wellbeing (Wu, 2016; Brady et al., 2017; Acemoglu, Ozdaglar, et al., 2023; Acemoglu, 2023). Relatedly, companies that have access to more data may possess an anticompetitive advantage relative to competitors, enabling them to exercise market power to extract surplus and to relax price competition in the marketplace, which would be detrimental for consumers (Acemoglu, 2024). Further, generative AI's capacity to activate linguistic patterns, including persuasion and rhetoric, could facilitate more tailored advertisement. The results would be particularly worrying if generative AI's communicative capabilities are combined with the data infrastructure of social media platforms to automate social engineering (McNealy, 2022), with potential uses in areas such as political communication.

An issue at the center of the debate is misinformation. Malicious actors can exploit generative AI to create false information in ways that convincingly copy the style and content of human-created text (Buchanan et al., 2021; Kreps et al., 2022; Lazer et al., 2018; Spitale et al., 2023), by synthetically generating text, audio, images, and videos ("deepfakes"). New malicious generative AI tools like WormGPT (a ChatGPT alternative for designing and refining cyber-attack strategies and malware; The Hacker News, 2023) or PoisonGPT (a modified open-source AI model designed to spread misinformation within a public data repository that is used to train other AI models; Mithril Security, 2023) show that these tools can be used to accomplish malign aims and to sabotage further technology development.

The possibility to create information that is personalized or targeted to specific individuals and groups is likely to increase as well, especially during elections (Benson, 2023). Politicians, including Republican presidential candidate Ron DeSantis, have already started using deepfakes in their political campaigns, such as fake images of Donald Trump hugging Anthoni Fauci (McCarthy, 2023). Manipulated political images make up a substantial portion (~20%) of visual misinformation on social media (Yang et al., 2023), and can be especially influential during elections and intergroup conflicts such as the Russo-Ukrainian and Israel-Gaza wars (Tworney et al., 2023). There is currently no legislation preventing the use of deepfakes in political campaigns. AI-assisted misinformation can spread rapidly on social media and research has already shown that micro-targeting people with deepfakes can influence their attitudes toward politicians (e.g., Dobber et al., 2021). Coupled with the fact that people are largely unable to tell the difference between AI- and human-generated text (Kreps et al., 2022) and that AI has been shown to generate more convincing misinformation than humans (Spitale et al., 2023), there are concerns that generative AI may also increase the quantity of misinformation. Indeed, hundreds of unreliable AI-news websites have popped up (Newsguard, 2023) and even ChatGPT can easily be prompted to generate misinformation.

This increase in misinformation may have significant social consequences. Political conspiracy theories and misinformation can affect voting decisions, health-related conspiracy theories can influence people's medical choices (e.g., vaccination), and misinformation and conspiracy theories can fuel conflict between groups (Douglas, 2021). While some people will simply ignore online misinformation (Acerbi et al., 2022), this content is likely to penetrate specific groups, especially since AI helps automate the micro-targeting process in which thousands of persuasive messages can now be generated easily at scale. For example, there is evidence that conservative voters were more susceptible to misinformation during the 2016 presidential election (Grinberg et al., 2019). Conspiracy theories and misinformation online can also contribute to attitude polarization (Del Vicario et al., 2016) and undermine trust (Tworney et al., 2023).

Therefore, regulation and interventions to limit the diffusion of AI-generated misinformation are needed. Simply warning people of deepfakes or including a tag clarifying whether a piece of content is AI- vs human-generated might backfire, as such tags have been shown to reduce the believability of true content as well (Longoni et al., 2022; Tworney et al., 2023). In the realm of human-generated misinformation on social media, psychological interventions based on accuracy-salience and educational interventions based on inoculation theory improve the quality of information shared. For example, making the concept of accuracy more salient can reduce the sharing of fake news, without adversely affecting the dissemination of accurate news (Pennycook et al., 2021; Pennycook and Rand, 2022). Moreover, endorsing accuracy not only decreases the sharing of false news but also increases the sharing of true news (Capraro and Celadin, 2023).

Inoculation theory or "prebunking" is a preemptive approach to countering misinformation that follows the vaccination analogy (Lu et al., 2023; McGuire, 1969; van der Linden, 2023). Several inoculation games and videos have been developed to expose subjects to controlled (i.e., weakened) doses of misinformation along with tools on how to spot it. For instance, in the Bad News Game, participants create fake news in a simulated social media setting with the aim to gather as many followers as possible. This activity makes them better at detecting online manipulation (Roozenbeek and Van der Linden, 2019; Traberg et al., 2022). Similarly, in a field study on YouTube, videos containing micro-doses of common misinformation techniques increased discernment of online manipulation tactics (Roozenbeek et al., 2022). Because prebunking conspiracy theories often works better than debunking them (Jolley and Douglas, 2017; Mason et al., 2023), future work could adapt these techniques to AI-generated news, especially in a way that mitigates cynicism of all (visual) media (Tworney et al., 2023).

One concern, however, is that interventions based on accuracy-salience and inoculation may be most effective for easily discernible misinformation. The risk that generative AI makes misinformation more subtle and harder to discern may necessitate a new toolbox of interventions, specifically designed to counteract (visual) AI-generated misinformation. Moreover, generative AI could lead to entirely new challenges, such as tackling misinformation disseminated via one-to-one personalized communications (e.g., through bots). This further highlights the urgency to adapt existing or develop a new set of intervention strategies (Feuerriegel et al., 2023).

Even in the absence of malicious actors, the most advanced AI-systems are known to “hallucinate” false information in a very realistic manner (Bubeck et al., 2023). These hallucinations may induce complex social dynamics, like self-fulfilling prophecies. Dating back to Merton (1948), a self-fulfilling prophecy is an initially false prediction that becomes true just because someone—e.g., a generative AI system—asserts that it will become true. In this sense, AI may produce prophecies that could “take a life for their own” (Citron and Pasquale, 2014). For example, automated scoring systems that predict the likelihood of default on debt repayment or of a job applicant being a bad hire may contribute to (or even cause) credit or employment risk.

Online communications are another important area of concern. Individuals may not know whether they are interacting with a person or a machine (Natale, 2021). This is increasingly likely to be the case when people engage online with businesses and with public services. If they believe, rightly or wrongly, that they are interacting with a machine, evidence suggests that their behavior is likely to change (March, 2021). Human behavior tends to become more selfish in human-machine interactions because reciprocation—a vital factor in sustaining cooperative, prosocial behavior—is not maintained as consistently as in verifiably human-human interactions (Ishowo-Oloko et al., 2019; Makovi et al., 2023). Generosity or cooperative behavior depends on people’s beliefs in each context about the relationship between the machine and the humans behind it (von Schenk et al., 2023). When interacting with a machine, people respond less emotionally, feeling less guilt about being ungenerous (Chugunova and Sele, 2022). They become more likely to be dishonest in pursuit of monetary rewards (Cohn et al., 2022). An outstanding research question concerns whether similar slippage from ethical standards occurs not only among people interacting with a machine, but also among those on the other side of the relationship who delegated to the machine (Köbis et al., 2021).

A critical but mostly overlooked implication is the impact of generative AI on the plurality of information available on the web. Companies including Microsoft and Google have envisioned integrating large language models such as ChatGPT and Bard into their proprietary search engines (recently, ChatGPT itself incorporated Bing search capabilities for some queries), but the implications of such a move have only started to be explored (Cutler, 2023). Among the most significant implications is users' access to information. Search engines that mobilize generative AI to provide tailored recommendations to users are likely to restrict the plurality of information available on the Web. When users input a query into the current version of Google Search, for instance, they are pointed to a plurality of sources. Although users tend to select among the first results returned by the search engine (Goldman, 2008), the interface enables them to browse a large number of alternative results. The same input directed to a search engine powered by generative AI will provide an extra layer of mediation that is likely to provide a much more limited amount of source information, unless specific design features are included to counteract this. The tool will still give users the impression of access to vast, nearly unlimited information that users customarily attributed to the Web, but it will actually reduce control over Web access to users, affecting their capacity to browse, explore and retrieve information available through the Web (Natale and Cooke, 2021).

In addition to reducing access to information, generative AI may also threaten the quality and availability of online information. The already-pervasive issue of bot accounts and autonomously generated content on social media and online message boards may be exacerbated by new generative technologies, which can assist in coding a multiplicity of these bots as well as providing text content for the bots to post (Ferrara et al., 2016; Yang and Menczer, 2023). These tools could also be used to generate content optimized for search engines en masse, a useful tactic for businesses to "poach" traffic from competitors' websites (e.g., Semrush, 2023). This is a problem because current generative AI tools essentially provide an "average" response to a particular question, and these models are trained largely on text data collected from the internet; therefore, if the practice of generating content optimized for search engines at massive scale becomes common practice, both generative AI tools and online information may crater into an average of averages, lacking true insight, creativity, or novel ideas. At the very least, it could make useful contributions difficult to identify within a sea of machine-generated content. Because companies often perform thought leadership or produce marketing content that doubles as an informational resource, the proliferation of unoriginal content could make it harder for novices to find the information that they need, raising new barriers to entry for those seeking to gain subject matter knowledge online.

An analogous concern is that common message boards and websites for knowledge sharing (e.g., Stack Overflow) have experienced both a reduction in questions posted—especially the sorts of basic questions that ChatGPT does well at answering—and an increase in question responses, perhaps due to writing aid from tools like ChatGPT (Burtch et al., 2023; Shan & Qiu, 2023; del Rio-Chanona et al., 2023). Though these Q&A sites require competent subject-matter experts to provide insights and suggestions, they also require neophytes to ask those questions in the first place. Reduced engagement by novice users not only has effects on the continued usefulness of these websites to aggregate and share knowledge, but also for innovation and creativity that may rely on content from these platforms as input. For example, some coding languages encourage user-written programs to expand the capabilities of the language; it is likely that entrepreneurial contributors are influenced in their decisions to create new programs based on common questions or demand from users on topic-specific Q&A sites. In the absence of questions being posted, user-contributions may not be optimally effective, since the user-feedback loops that developers rely upon are interrupted by the information provisional capabilities of generative AI platforms. In net, this could forestall newcomers from acquiring specialized knowledge and technical skills, in addition to preventing the development of other access-equalizing tools.

In conclusion, while generative AI has the potential to expand access to and content of information, it also raises significant challenges such as market anticompetitive advantages, data misuse, data poisoning, misinformation proliferation, and altered human-machine interactions, all of which necessitate careful consideration and targeted research. Table 4 summarizes the main directions of future research. The next section discusses policy recommendations (cf. Table 5).

Future research directions
Understand how the largest firms could monopolize the future of AI; find ways for smaller and innovative firms to effectively compete with those largest players. Investigate strategies to identify and limit the spread of misinformation generated by AI. Explore regulatory measures to prevent misuse or inappropriate access to data by AI systems. Investigate how AI can be used to make information more accessible, especially for individuals with disabilities. Explore ways to design AI-systems that support ethical behavior in human-machine interactions. Examine how AI-enhanced search engines can be designed to preserve user autonomy and plurality of information. Consider how the proliferation of AI-generated content could lower the quality of online information and ensure that human users can continue to contribute new knowledge.

Table 4. Summary of the main research directions on the impact of generative AI on information.

Policymaking in the age of artificial intelligence

Regulation of AI

The rapid popularization of generative AI models has prompted many governments worldwide to begin building regulatory frameworks. The challenges raised by generative AI are global in nature (Jobin et al., 2019). However, the responses to these challenges so far have been specific to individual countries or areas. In this article, we focus on the regulatory responses of the EU, US, and UK. Regulations are also being developed in other major countries, including China and India (Haridas et al., 2023; Roberts, 2023).

The European Union's AI Act has emerged as one of the first major attempts to provide a legal framework for the development and deployment of AI (European Parliament, 2023a). The Act aims to address the challenges posed by AI technologies while fostering innovation and trust in AI applications (European Parliament, 2023b). This initiative comes with several pros. Firstly, it introduces a risk-based regulatory approach, distinguishing between high-risk and low-risk AI applications. This categorization ensures that AI systems with significant implications for individuals' rights and safety are subject to stricter scrutiny and compliance requirements. Secondly, the Act emphasizes transparency and accountability in AI systems, requiring clear information about how AI decisions are made, particularly in high-risk scenarios. Additionally, the Act promotes ethical AI development, focusing on fundamental rights, non-discrimination, and privacy. However, the Act is not without its cons (Morgan Lewis, 2023). The broad definitions and categories within the Act pose challenges, creating potential uncertainty for AI developers and users. Further, the strict regulations might place EU companies at a competitive disadvantage globally, particularly against firms in regions with more lenient AI laws.

In contrast, the US has historically had a more fragmented approach, with various federal and state-level initiatives rather than a single, comprehensive legislative framework (Felz et al., 2022). This approach has its advantages. For one, it allows for more flexibility and adaptability in regulation, catering to the diverse range of AI applications and industries in the US. It also promotes a more innovation-friendly environment by avoiding overly prescriptive rules that could hinder technological advancement. However, the US approach also has notable disadvantages. The lack of a unified regulatory framework can lead to inconsistencies and uncertainties, potentially creating a complex patchwork of regulations for AI companies to navigate. This fragmented approach might also lag in addressing broader ethical and social concerns about AI, such as privacy, bias, and accountability. Further, without a cohesive national strategy, the US risks falling behind in setting global standards for AI governance. On October 30, 2023, President Biden issued an Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, which directs the development of new guidelines, reports, and governance structures relating to AI, representing an effort to establish a more cohesive federal policy on AI (White House, 2023).

In the UK, the government has published a White Paper advocating for a pro-innovation approach, particularly in commercial applications of AI (Department for Science Innovation and Technology, 2023). While the White Paper recognizes the risks of AI and the challenge of building public trust, it refrains from proposing a regulatory framework to encourage innovation, which contrasts with the EU's approach. Instead, the White Paper outlines some "cross-sectoral" non-statutory soft principles: safety, security, robustness, appropriate transparency and explainability, fairness, accountability and governance, and contestability and redress. The White Paper opts against a specialist AI regulator, preferring to support existing regulators in integrating AI considerations. Furthermore, the focus on commercial innovations has drawn criticism for overlooking the increasing use of AI in government sectors like healthcare and education. One leading non-governmental organization, the Public Law Project, led a civil society coalition to produce *Key principles for an alternative AI white paper* (Public Law Project, 2023), which argues that an alternative vision is necessary as the "white paper, misses a vital opportunity to ensure that fundamental rights and democratic values are protected… it fails to ensure that adequate safeguards and standards are in place for use of AI by public authorities." Amongst other proposals, the alternative white paper argues that: government use of AI must be transparent, transparency requirements must be mandatory, there must be clear mechanisms for accountability, the public should be consulted about new automated decision-making tools before they are deployed by government, there must be a specialist regulator to enforce the regulatory regime and ensure people can seek redress when things go wrong, and uses of AI that threaten fundamental rights should be prohibited.

In sum, the regulations of the EU, the US, and the UK do not pay sufficient attention to socioeconomic inequalities. In the following, we outline several key interventions currently missing from these regulatory frameworks (Acemoglu, Autor et al., 2023). See Table 5 for a summary.

Tax system: Current tax codes in many developed countries often place a heavier burden on firms that hire labor than on those that invest in algorithms to automate work (Abbott and Bogenschneider, 2018; Acemoglu et al., 2020). This has resulted in a lower share of income to labor while capital investments are rewarded. We should aim to create a more symmetric tax structure, where marginal taxes for hiring (and training) labor and for investing in equipment/software are equated. This will help to shift incentives toward human-complementary technological choices by reducing the bias of the tax code toward physical capital over human capital.

Labor voice and control of consumer information: Given that AI will have tremendous impact across industries and throughout society, it would be prudent to ensure that workers and civil society have a voice in this change. Health and safety rules should also be updated accordingly. In addition, data unions could be helpful to put the power and benefits of user data back in the hands of consumers. Given the concerns that a handful of very large companies will control the direction of generative AI, it makes sense that users be compensated for the use of their information, or enabled to support other emergent competitors to predominant market players like Microsoft and Google.

Funding for more human-complementary research: Because the current path of research is biased toward automation (Acemoglu and Restrepo, 2019; Autor et al., 2022; Acemoglu and Johnson, 2023), support for research and development of human-complementary AI technologies could offer strong upsides for growth. It is most feasible to focus on specific sectors and activities where opportunities are already abundant. These include education, healthcare, and modern craft worker training—where the information provisional capabilities of AI systems could boost productivity and enable workers to earn higher wages by augmenting their skills. In the US, DARPA orchestrated investments and competitions to foster development of self-driving cars and dexterous robotics—in a similar fashion, governments should encourage competition and investment that pairs AI tools with human expertise, aiming to improve work in vital social sectors.

Professional development and training: Investment in professional development and training is crucial for professionals such as educators and healthcare workers to effectively integrate AI tools into their work. Training programs should focus on the capabilities and limitations of AI, include ethical considerations, and teach technical skills required to interact with AI systems. Such training will empower professionals to use AI as a complementary tool that enhances their skills.

Combating AI-generated misinformation: Given the substantial impact that generative AI can have on the quality and quantity of misinformation circulated online, especially in sensitive areas such as political campaigns and news media, it is critical for governments to invest in combating AI-generated misinformation. Tools and standards to identify AI-generated content, including text, images, audio, and video, should be developed. Additionally, educational campaigns should be initiated, to reduce general susceptibility to misinformation and provide the informed public with (currently deficient) fact-checking strategies. A task force composed of policymakers, technology companies, and social scientists could help develop practical methods to effectively combat a potential infodemic.

Governmental and consultative expertise: To foster human-complementary AI integration, it is fundamental to have AI expertise within the government. AI will touch every area of government investment, regulation, and oversight, including transportation, energy production, labor, healthcare, education, environmental protection, public safety, and military capacity. Developing consultative AI bodies that can advise governments and support the many agencies and regulators tackling these challenges will support more timely and effective decision-making.

Policy recommendations
Create a more symmetric tax structure, where marginal taxes for hiring and training labor and for investing in the development, installation and usage of new AI tools are equated. Involve workers and civil society in AI-related changes and establish data unions to empower data owners with control over their data. Increase support for research into human-complementary AI tools to enhance productivity and workers’ skillsets. Train professionals, especially in education and healthcare, in the use of AI tools, covering their capabilities, limitations, and ethical considerations. Invest in strategies aimed to combat AI-generated misinformation, including developing tools that can identify AI-generated misinformation and initiating educational campaigns. Embed AI expertise within government to advise and support decision-making across various sectors.

Table 5. Policy recommendations for mitigating socioeconomic inequalities potentially caused by generative AI in the workplace, education, healthcare, and information, and not covered by current regulatory approach in the EU, US, and UK.

Regulation using AI

Generative AI holds enormous potential to provide policy suggestions, due to its capacity to analyze vast amounts of data, recognize complex patterns, and offer insights that might elude human analysis. Such analysis can uncover hidden relationships and forecast future trends, providing a data-driven foundation for policy decisions. Moreover, AI’s potential to simulate various policy outcomes based on historical data and predictive models can aid policymakers in understanding the potential impacts of their decisions (Pencheva et al., 2020; Zuiderwijk et al., 2021; Madan and Ashok, 2023). AI could also automate some aspects of the policymaking process itself, as recently demonstrated in the context of tax policy design (Zheng et al., 2022).

Integrating generative AI into policymaking processes could have the potential to overcome human biases and limitations, but it could also perpetuate these biases. In general, the ethical and practical concerns of using AI for policymaking are significant and perhaps even prohibitive with the current tools available.

Generative AI used in policymaking needs to align with human values, but this is far from trivial. Humans have a wide range of culturally diverse beliefs about right and wrong, and aligning AI systems to human values and preferences is challenging even within narrow domains such as automated driving (Awad et al., 2018). Aligning generative AI, especially in the domain of policy recommendations, becomes even more challenging. For example, consider different normative principles that have been identified for trustworthy ethical AI (Floridi and Cowls, 2019). It is argued that AI should, amongst other things, promote beneficence (promote human well-being and welfare); non-maleficence (not cause harm and generate outputs that assist in carrying out illegal, harmful, or immoral actions); justice (preserve fairness, justice, and solidarity: it should not generate outputs that discriminate against certain groups, especially marginalized groups), and ensure autonomy (respecting human freedom and ensuring humans should choose how and whether to delegate policy decisions). While these principles are all defensible in the abstract—forming the basis of much normative ethical theory and applied ethics—challenges will inevitably arise when these principles conflict. Generative AI may assist in generating policies that maximize overall aggregate welfare, in line with utilitarian philosophy (Mill, 1861) but in doing so infringe on human rights (neglecting the principle of autonomy) or recommending some harm to a smaller group for the benefit of the majority (neglecting the principle of non-maleficence). There is no consensus amongst laypeople about how such moral dilemmas should be resolved (Greene et al., 2001; Kahane et al., 2018), nor is there normative agreement amongst philosophers on how they should be resolved and why. This discord among human thinkers underscores the challenge of programming AI to make policy decisions that involve moral trade-offs.

To make an explicit example, consider the following three high-level, high-priority constraints for aligned chatbots: 1) to not cause harm or provide dangerous information; 2) to not generate outputs that discriminate against certain groups, and 3) to be culturally sensitive. While these objectives are all desirable, they are increasingly difficult to reconcile. If the only constraint is to avoid dangerous information (non-maleficence), regardless of social neutrality or cultural sensitivity (justice), one can use reinforcement learning from human feedback (Ouyang et al., 2022) using a convenience sample of annotators. But this approach would fail to ensure social neutrality, since a convenience sample of annotators would have unrepresentative biased views on what constitutes immoral actions or generally undesirable outputs (Peters, 2022; Santurkar et al., 2023). Political and social neutrality may be approximated by engaging in carefully balanced reinforcement learning from human feedback, based on a broadly representative array of opinion, or by having a singular chatbot that facilitates consensus-making among diverse human values (Bakker et al., 2022). Alternatively, an ecosystem of chatbots with diverse systems of values—liberal and conservative bots, secular and religious bots, etc.—may emerge. These chatbots can each focus on their specific domain, while also undergoing a political process to achieve collective decisions among themselves. However, these approaches would still fail at cultural sensitivity, since different cultures may be different in terms of the social groups they include, the topics these groups value, and the range of these cultural values (Atari et al., 2023).

In the worst case scenario, then, the alignment of generative AI would be entirely based on the views of a small group of socially, politically, and culturally homogeneous informants (Santy et al., 2023; Acemoglu and Johnson, 2023). But even in the best-case scenario, where generative AI is trained on a diverse and nuanced set of preferences, we would still have significant problems. Even if we have a more diverse set of information about humans' actual values and how they might want trade-offs to be made for moral dilemmas, we still lack widespread agreement on a specific normative standard to justify these descriptive preferences.

Besides the alignment problem, there is the implementation problem: how to equip policymakers with reliable support from organizations and specialized staff. Most policymakers currently lack the knowledge and skills to directly evaluate the extent to which AI-based generative chatbots may embed undesired preferences or detrimental systematic biases. Admittedly, one can hardly expect that policymakers can acquire the needed knowledge and skills in due time. So, policymakers are likely to become the "principals" in a principal-agent problem, struggling to take into account the preferences of their "AI-agents". Policymakers will have to rely on some other agent for this evaluation, based on scientific principles for characterizing machine behavior and misbehavior (Rahwan et al., 2018).

Hence, it is crucial to design supporting organizations that systematically provide the policymakers with: (i) frequent evaluation of the current state of alignment between legal and regulatory requirements on one hand, and the behaviors exhibited by AI-based chatbots, and (ii) mechanisms to signal any legal and regulatory changes in those requirements to companies that operate AI-based chatbots—thus putting *society in the loop* (Rahwan, 2018). These desiderata in turn require the construction of a dedicated office in the organization that monitors the AI-based chatbot, taking into account—and possibly predicting—its evolution.

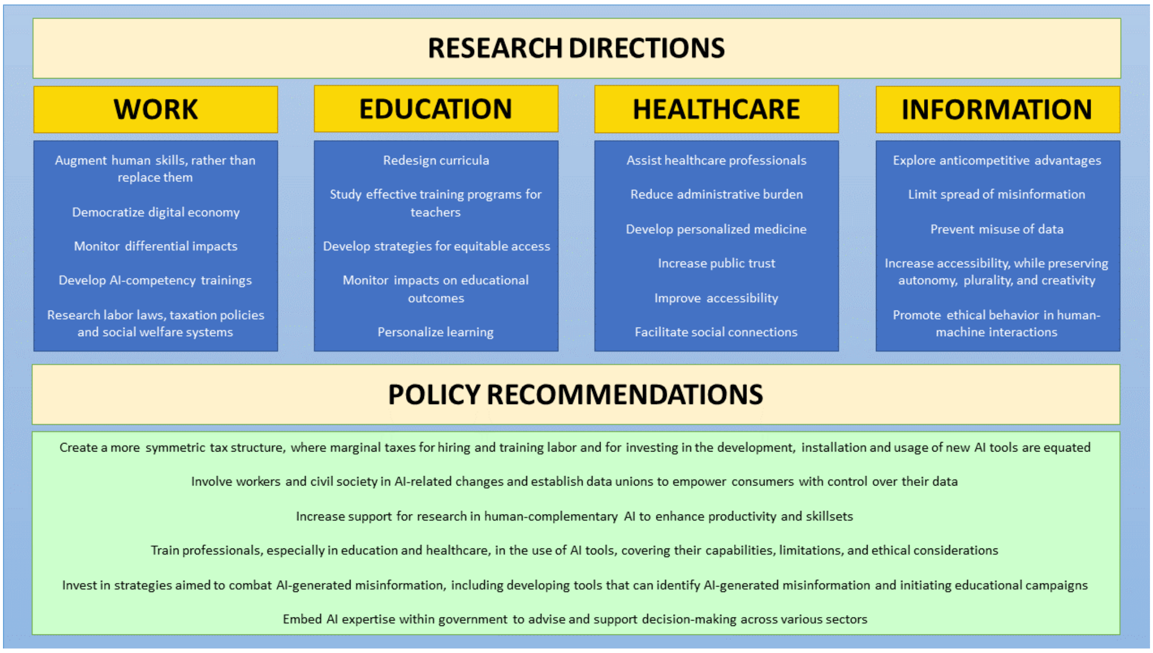
Finally, there is also a philosophical problem. Even if we could solve the problems of conflicting preferences, even if we could generate a good culturally sensitive sample, and even if we could solve the implementation problem - should we?

Conclusion

The future will likely be starkly different from anything we have experienced before. But the impact of generative AI is neither inherently positive nor negative. The effects of generative AI will ultimately depend on the choices that we make to design and deploy the technology. We stand at a unique and historical moment; our decisions and actions today will demonstrably shape the trajectory of our future. This responsibility extends to all sectors of society, including governance, scientific research, industry, and the general public.

In this article, we have specifically focused on the socioeconomic inequalities that are likely to be impacted by the advent of generative AI. This technology has profound implications for critical domains such as work, education, health, and information. For instance, in the workplace, AI could automate some job tasks, create new work, change wage distributions, and require new skill sets. In education, AI could democratize learning and provide personalized education solutions, but also raises concerns about the digital divide. In the healthcare sector, AI's ability to analyze large datasets can lead to better patient outcomes, but it also raises questions about equitable access to AI-driven healthcare services and the irreplaceable value of human interactions. In the domain of information, AI has the potential to offer more tailored, efficient, and democratic ways to process information, yet it also poses challenges related to misinformation and diversity of thought.

We have outlined several research questions that urgently require answers to address these issues effectively. These questions aim to harness AI's benefits while mitigating its risks. Additionally, we have observed that current regulatory approaches in the European Union, the United States, and the United Kingdom sometimes fail to adequately address these emerging challenges. There is a need for a dynamic regulatory framework that can keep pace with the rapid advancements in AI technology. See Figure 1 for an infographic summarizing the main points of the article.



RESEARCH DIRECTIONS

WORK	EDUCATION	HEALTHCARE	INFORMATION
Augment human skills, rather than replace them Democratize digital economy Monitor differential impacts Develop AI-competency trainings Research labor laws, taxation policies and social welfare systems	Redesign curricula Study effective training programs for teachers Develop strategies for equitable access Monitor impacts on educational outcomes Personalize learning	Assist healthcare professionals Reduce administrative burden Develop personalized medicine Increase public trust Improve accessibility Facilitate social connections	Explore anticompetitive advantages Limit spread of misinformation Prevent misuse of data Increase accessibility, while preserving autonomy, plurality, and creativity Promote ethical behavior in human-machine interactions

POLICY RECOMMENDATIONS

Create a more symmetric tax structure, where marginal taxes for hiring and training labor and for investing in the development, installation and usage of new AI tools are equated

Involve workers and civil society in AI-related changes and establish data unions to empower consumers with control over their data

Increase support for research in human-complementary AI to enhance productivity and skillsets

Train professionals, especially in education and healthcare, in the use of AI tools, covering their capabilities, limitations, and ethical considerations

Invest in strategies aimed to combat AI-generated misinformation, including developing tools that can identify AI-generated misinformation and initiating educational campaigns

Embed AI expertise within government to advise and support decision-making across various sectors

Figure 1. Infographic that summarizes the main research directions and policy recommendations suggested in the article.

Our hope is that this work contributes to developing a comprehensive research agenda and to sparking public debates on these critical topics. As we stand at the cusp of this new era, it is crucial that we engage in thoughtful and inclusive discussions about the role of AI in shaping our society, because the decisions we make today will have lasting impacts on generations to come.

References

Abbott, R., & Bogenschneider, B. (2018). Should Robots Pay Taxes? Tax Policy in the Age of Automation. *Harvard Law & Policy Review, 12*, 145-175.

Abdou, C. M., & Fingerhut, A. W. (2014). Stereotype threat among black and white women in healthcare settings. *Culturally Diverse Ethnic Minority Psychology, 20,* 316–323.

Acemoglu, D. (2023). Written testimony for hearing on "The Philosophy of AI: Learning from History, Shaping Our Future." *Senate Committee on Homeland Security and Governmental Affairs.* Retrieved from https://www.hsgac.senate.gov/wp-content/uploads/Testimony-Acemoglu-2023-11-08.pdf

Acemoglu, D. (2024). Harms of AI. In J. Bullock (Ed.), *The Oxford Handbook of AI Governance,* forthcoming. Oxford University Press. Retrieved December 4, 2023, from https://academic-oup-com.libproxy.mit.edu/edited-volume/41989/chapter/411053764

Acemoglu, D., Ahmed, F., Hart, A.J., & Johnson, S. (2023). From Automation to Augmentation: Redefining Engineering Design and Manufacturing in the Age of NextGen AI. *In progress.*

Acemoglu, D., Autor, D., Dorn, D., Hanson, G., & Price, B. (2016). Import competition and the great U.S. employment sag of the 2000s. *Journal of Labor Economics, 34,* 141-198.

Acemoglu, D., Autor, D., & Johnson, S. (2023). Can we have pro-worker AI? MIT Shaping the Future of Work Initiative, policy memo. Retrieved from https://shapingwork.mit.edu/wp-content/uploads/2023/09/Pro-Worker-AI-Policy-Memo.pdf

Acemoglu, D., & Johnson, S. (2023). *Power and progress: Our 1000-year struggle over technology and prosperity*. PublicAffairs, Hachette.

Acemoglu, D., Manera, A., & Restrepo, P. (2020). Does the U.S. tax code favor automation? *Brookings Papers on Economic Activity.* Retrieved December 7, 2023, from https://www.brookings.edu/articles/does-the-u-s-tax-code-favor-automation/

Acemoglu, D., Ozdaglar, A., & Siderius, J. (2023). A Model of Online Misinformation. *Review of Economic Studies*, forthcoming.

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, *108*, 1488-1542.

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives, 33*, 3-30.

Acemoglu, D., & Restrepo, P. (2022a). Tasks, automation, and the rise in U.S. wage inequality. *Econometrica, 90*, 1973-2016.

Acemoglu, D., & Restrepo, P. (2022b). Demographics and automation. *Review of Economic Studies, 89*, 1-44.

Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School (HKS) Misinformation Review, 3.*

Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. NBER Working Paper No. 31422.

Akgun, S. & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI Ethics, 2,* 431–440.

Aksoy, C. G., Barrero, J. M., Bloom, N., Davis, S. J., Dolls, M., & Zarate, P. (2023). *Working from home around the globe: 2023 Report* (No. 53). EconPol Policy Brief.

Andreessen, M. (2023). The Techno-Optimist Manifesto. *Andreessen Horowitz.* https://a16z.com/the-techno-optimist-manifesto/

Aristidou, A., Jena, R., & Topol, E. J. (2022). Bridging the chasm between AI and clinical implementation. *Lancet, 399*, 620.

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which Humans? *Available at: https://doi.org/10.31234/osf.io/5b26t*

Autor, D. (2019). Work of the Past, Work of the Future. *AEA Papers and Proceedings, 109*, 1–32.

Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2022). New frontiers: The origins and content of new work, 1940–2018. *NBER Working Paper no. 30389*.

Autor, D., Katz, L., & Krueger, A. (1998). Computing inequality: Have computers changed the labor market? *Quarterly Journal of Economics, 113,* 1169–1213.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*, 59-64.

Ayers, J. W. et al (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine, 183*, 589-596.

Baker, R. & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32,* 1052–1092.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., ... & Summerfield, C. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems, 35*, 38176-38189.

Beckes, L., & Sbarra, D. A. (2022). Social baseline theory: State of the science and new directions. *Current Opinion in Psychology, 43,* 36-41.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Benson, T. (2023). This disinformation is just for you. *Wired*. https://www.wired.com/story/generative-ai-custom-disinformation/

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. From *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, ed. Smit, I., et al. Institute of Advanced Studies in Systems Research and Cybernetics. 2:12–17.

Brady, W., Wills, J., Jost, J., Tucker, J., & Van Bavel, J. (2017). Emotion Shapes the Diffusion of Moralized Content in Social Networks. *Proceedings of the National Academy of Sciences, 114,* 7313–7318.

Breakstone, J., Smith, M., Wineburg, S., Rapaport, A., Carle, J., Garland, M., & Saavedra, A. (2021). Students' civic online reasoning: A national portrait. *Educational Researcher, 50,* 505-515.

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at work. *NBER Working Paper No. 31161*.

Brynjolfsson, E., & McAfee, A. (2016). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W.W. Norton.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *Available at: https://arxiv.org/abs/2303.12712*.

Buchanan, B., Lohn, A., Musser, M., & Sedova, K. (2021). Truth, lies, and automation: How language models could change disinformation. *Center for Security and Emerging technology*. https://cset.georgetown.edu/publication/truth-lies-and-automation/

Burtch, G., Lee, D., & Chen, Z. (2023). The Consequences of Generative AI for UGC and Online Community Engagement. Available at *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4521754*

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*, 1636-1642.

Campbell, J. (2023). AI's future: Utopia or dystopia? Experts weigh in on five possible outcomes. *AI News Today. https://ainewstoday.co.uk/2023/05/07/ais-future-utopia-or-dystopia-experts-weigh-in-on-five-possible-outcomes/*

Capraro, V., & Celadin, T. (2023). "I think this news is accurate": Endorsing accuracy decreases the sharing of fake news and increases the sharing of real news. *Personality and Social Psychology Bulletin*, *49*, 1635-1645.

Cené, C. W., et al. (2022). Effects of Objective and Perceived Social Isolation on Cardiovascular and Brain Health: A Scientific Statement From the American Heart Association. *Journal of the American Heart Association, 11*, e026493.

Citron, D. K. & Pasquale, F. A. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*, 1-33.

Chan, C. K. Y. & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-023-00411-8

Chiu, T. K. F. (2023). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*.

Chugunova, M., & Sele, D. (2022). We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics,* 99, 101897.

Cohn, A., Gesche, T., & Maréchal, M. A. (2022). Honesty in the digital age. *Management Science*, 68, 827-845.

Counts, S., Suri, S., Brown, A., Xu, B., R Raghavan, S. (2022). Who gets to work in the digital economy? *Business and Society*. Available at: https://hbr.org/2022/08/who-gets-to-work-in-the-digital-economy

Cuevas, A. G., O’Brien, K., & Saha, S. (2016). African American experiences in healthcare: “I always feel like I’m getting skipped over”. *Health Psychology, 35*, 987–995.

Cutler, K. (2023). ChatGPT and search engine optimisation: The future is here. *Applied Marketing Analytics, 9,* 8-22.

Del Carmen, M. G., Herman, J., Rao, S., Hidrue, M. K., Ting, D., Lehrhoff, S. R., ... & Ferris, T. G. (2019). Trends and factors associated with physician burnout at a multispecialty academic faculty practice organization. *JAMA Network Open*, *2*, e190554-e190554.

del Rio-Chanona, M., Laurentsyeva, N., & Wachs, J. (2023). Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow. *Available at https://arxiv.org/abs/2307.07367*

Dell’Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Department for Science, Innovation, and Technology (2023). AI-regulation: A pro-innovation approach. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

Del Vicario, M., Bessi, A., Zollo, F., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences, 113*. 554-559.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes?. *The International Journal of Press/Politics*, *26*, 69-91.

Douglas, K. M. (2021). Are conspiracy theories harmless? *The Spanish Journal of Psychology*, *21*, e13.

Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, *47*, 459-469.

Dovidio, J. F., Penner, L. A., Albrecht, T. L., Norton, W. E., Gaertner, S. L., & Shelton, J. N. (2008). Disparities and distrust: The implications of psychological processes for understanding racial disparities in health and health care. *Social Sciences and Medical Journal*, *67*, 478–486.

Eling, M., Nuessle, D. & Staubli, J. (2022). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. *The Geneva Papers on Risk and Insurance: Issues and Practice, 47*, 205–241.

Elsen-Rooney, M. (2023). NYC education department blocks ChatGPT on school devices, networks. https://www.chalkbeat.org/newyork/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence/

Engelbart, D. C. (1995). Toward augmenting the human intellect and boosting our collective IQ. *Communications of the ACM*, *38*, 30-32.

European Parliament (2023a). https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

European Parliament (2023b). https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

Felz, D. J., Peretti, K. K., & Austin, A. (2022). Privacy, cyber and data strategy advisor: AI regulation in the U.S.: What's coming, and what companies need to do in 2023. https://www.alston.com/en/insights/publications/2022/12/ai-regulation-in-the-us

Ferrara, E., Varol, O. Davis, C., Menczer, F. & Flammini, A. (2016). The rise of social bots. *Communications of the Association for Computing Machinery*, 59(7): 96–104.

Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., Lorenz-Spreen, P., Tomz, M., & Pröllochs, N. (2023). Research can help to tackle AI-generated disinformation. *Nature Human Behaviour*, 1-4.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8

Goldin, C., & Katz, L. (2008). "The evolution of U.S. educational wage differentials, 1890 to 2005." *The race between education and technology*, Harvard University Press. Chapter 8.

Goldman, E. (2008). Search engine bias and the demise of search engine utopianism. In: Spink A and Zimmer M (eds) *Web Search: Multidisciplinary Perspectives*. Berlin: Springer, pp.121–133.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105-2108.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*, 374–378.

Haridas, G., Sohee, S. K., & Brahmecha, A. (2023). The key policy frameworks governing AI in India. *https://accesspartnership.com/the-key-policy-frameworks-governing-ai-in-india/*

Hoes, E., Altay, S., & Bermeo, J. (2023). Leveraging ChatGPT for efficient fact-checking. Available at https://osf.io/preprints/psyarxiv/qnjkf/

Holt-Lunstad, J. (2021). The Major Health Implications of Social Connection. *Current Directions in Psychological Science, 30*, 251-259.

Holt-Lunstad, J. (2022). Social Connection as a Public Health Issue: The Evidence and a Systemic Framework for Prioritizing the "Social" in Social Determinants of Health. *Annual Review of Public Health, 43*, 193-213.

Holt-Lunstad, J., Smith, T. B., Layton, J. B. (2010). Social Relationships and Mortality Risk: A Meta-analytic Review. *PLOS Medicine, 7,* e1000316.

Hommel, K., Madsen, M., & Kamper, A. L. (2012). The importance of early referral for the treatment of chronic kidney disease: A Danish nationwide cohort study. *BMC Nephrology, 13*, 108–116.

Ishowo-Oloko, F., Bonnefon, J. F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence, 1*, 517-521.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389-399.

Johnson, M. (2021). A scalable approach to reducing gender bias in Google translate. https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html

Johnson, M., Albizri, A. & Harfouche, A. (2021). Responsible Artificial Intelligence in Healthcare: Predicting and Preventing Insurance Claim Denials for Economic and Social Wellbeing. *Information Systems Frontiers*, *25*, 2179-2195.

Johnson, B. S., *et al.* (2023). Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectrum*, *7,* pkad015.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*, 131-164.

Kanjee, Z., Crowe, B., & Rodman, A. (2023). Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA*, 330, 78–80.

Kasneci, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Köbis, N., Bonnefon, J. F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, 5, 679-685.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61: 1281–1317.

Kranzberg, M. (1985). The information age: Evolution or revolution. *Information technologies and social transformation*, 35-54.

Kreps, S., McCain, R., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science, 9*, 104-117.

Kumar, A., & Epley, N. (2021). It's surprisingly nice to hear you: Misunderstanding the impact of communication media can lead to suboptimal choices of how to connect with others. *Journal of Experimental Psychology: General*, *150*, 595-607.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*, 1094-1096.

Leigh-Hunt, N., et al. (2017). An overview of systematic reviews on the public health consequences of social isolation and loneliness. *Public Health, 152*, 157-171.

Licklider, J. C. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, 4-11.

Liu, M., et al. (2023). Future of education in the era of generative artificial intelligence: Consensus among Chinese scholars on applications of ChatGPT in schools. *Future Education Research, 1,* 72–101.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*, 629-650.

Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022). News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 97-106).

López-Cevallos, D. F., Harvey, S. M., & Warren, J. T. (2014). Medical mistrust, perceived discrimination, and satisfaction with health care among young-adult rural Latinos. *The Journal of Rural Health*, *30*, 344–351.

Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X. D. (2023). Psychological Inoculation for Credibility Assessment, Sharing Intention, and Discernment of Misinformation: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *25,* e49255.

Madan, R., & Ashok, M. (2023). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly*, *40*, 101774.

Makovi, K., Sargsyan, A., Li, W., Bonnefon, J. F., & Rahwan, T. (2023). Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nature Communications, 14,* 3108.

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, 87, 102426.

Mason, A. M., Compton, J., Tice, E., Peterson, B., Lewis, I., Glenn, T., & Combs, T. (2023). Analyzing the Prophylactic and Therapeutic Role of Inoculation to Facilitate Resistance to Conspiracy Theory Beliefs. *Communication Reports*, 1-15.

McCarthy, B. (June 7, 2023). Ron DeSantis ad uses AI-generated photos of Trump, Fauci. *AFP*. *https://factcheck.afp.com/doc.afp.com.33H928Z*

McGuire, W. J. (1964). Some contemporary approaches. In *Advances in experimental social psychology* (Vol. 1, pp. 191-229). Academic Press.

McGrew, S. (2024). Teaching Lateral Reading: Interventions to Help People Read like Fact Checkers. *Current Opinion in Psychology*, *55*, 101737.

McNealy, J. E. (2022). Platforms as phish farms: Deceptive social engineering at scale. *New Media & Society*, *24*, 1677–1694.

Merton, R. K. (1948). The Self-Fulfilling Prophecy. *The Antioch Review*, *8*, 193-210.

Mill, J. S. (1861/2016). Utilitarianism. In *Seven masterpieces of philosophy* (pp. 329-375). Routledge.

Mithril Security (2023). PoisonGPT: How we hid a lobotomized LLM on hugging face to spread fake news. https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/

Morewedge, C. K., Mullainathan, S., Naushan, H. F., Sunstein, C. R., Kleinberg, J., Raghavan, M., & Ludwig, J. O. (2023). Human bias in algorithm design. *Nature Human Behaviour*, *7*, 1822–1824.

Morgan Lewis (2023). https://www.morganlewis.com/pubs/2023/10/european-trilogue-session-on-eu-ai-act-concludes-with-questions-remaining#:~:text=October%2026%2C%202023,AI%20Act%20in%20May%202023

Natale, S. (2021). *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test*. Oxford University Press.

Natale, S., & Cooke, H. (2021). Browsing with Alexa: Interrogating the impact of voice assistants as web interfaces. *Media, Culture & Society*, 43, 1000-1016.

Nov, O., Singh, N., & Mann, D. M. (2023). Putting ChatGPT's medical advice to the (Turing) Test. *Available at https://www.medrxiv.org/content/10.1101/2023.01.23.23284735v2*

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*, 187–192.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730-27744.

Pasquale, F. (2020). *New laws of robotics*. Harvard University Press.

Patel, S. B., & Lam, K. (2023). ChatGPT: the future of discharge summaries? *Lancet Digit Health*, 5, e107–e108.

Pencheva, I., Esteve, M., & Mikhaylov, S. J. (2020). Big Data and AI–A transformational shift for government: So, what next for research?. *Public Policy and Administration*, *35*, 24-44.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *Available at https://arxiv.org/abs/2302.06590.*

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590-595.

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*, 2333.

Penninkilampi, R., Casey, A. N., Singh, M. F., & Brodaty, H. (2018). The association between social engagement, loneliness, and risk of dementia: a systematic review and meta-analysis. *Journal of Alzheimer's Disease, 66*, 1619-1633.

Peters, U. (2022). Algorithmic political bias in artificial intelligence systems. *Philosophy & Technology, 35*, 25.

Promberger, M., & Baron, J. (2006). Do patients trust computers?. *Journal of Behavioral Decision Making*, *19*, 455-468.

Public Law Project (2023). Key principles for an alternative AI white paper. https://publiclawproject.org.uk/content/uploads/2023/06/AI-alternative-white-paper-in-template.pdf

Quinn, T. P., Senadeera, M., Jacobs, S., Coghlan, S., & Le, V. (2021). Trust and medical AI: the challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association*, *28*, 890-894.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology, 20*, 5-14.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature, 568*, 477-486.

Rathbone, J. A., Cruwys, T., Jetten, J., & Barlow, F. K. (2020). When stigma is the norm: How weight and social norms influence the healthcare we receive. *Journal of Applied Social Psychology, 53*, 185-201.

Ren, F., Zhou, Y. (2020). CGMVQA: A new classification and generative model for medical visual question answering. *IEEE Access, 8*, 50626–50636.

Restrepo, P. (2023). Automation: Theory, Evidence, and Outlook. *NBER Working Paper No. 31910*.

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., & Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*, 14952.

Roberts, H. (2023). The future of AI policy in China. *East Asia Forum. https://www.eastasiaforum.org/2023/09/27/the-future-of-ai-policy-in-china/*

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*, 1-10.

Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8,* eabo6254.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Available at https://arxiv.org/abs/2303.17548*.

Santy, S., Liang, J. T., Bras, R. L., Reinecke, K., & Sap, M. (2023). NLPositionality: Characterizing Design Biases of Datasets and Models. *Available at https://arxiv.org/abs/2306.01943.*

Selwyn, N., Hillman, T., Bergviken Rensfeldt, A., & Perrotta, C. (2023). Digital technologies and the automation of education. *Postdigital Science and Education*, *5*, 15-24

Semrush Team. (2023). Maximizing SEO Impact with ChatGPT: A Comprehensive Guide. *Semrush Blog*. https://www.semrush.com/blog/chatgpt-seo/

Septiandri, A.A., Constantinides, M., & Quercia, D. (2023). The impact of AI innovations on U.S. occupations. *Nokia Bell Labs, Cambridge, UK*, work in progress.

Shan, G., & Qiu, L. (2023). Examining the Impact of Generative AI on Users' Voluntary Knowledge Contribution: Evidence from A Natural Experiment on Stack Overflow. *Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4462976*

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *Science, 9(26)*, eadh1850.

Stoyanovich, J., Van Bavel, J. J., & West, T. V. (2020). The imperative of interpretable machines. *Nature Machine Intelligence*, *2*, 197-199.

The Hacker News (2023). WormGPT: New AI tool allows cybercriminals to launch sophisticated cyber attacks. https://thehackernews.com/2023/07/wormgpt-new-ai-tool-allows.html

The Lancet Regional Health – Europe, Embracing generative AI in health care (2023). *The Lancet Regional Health – Europe, 30.*

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *2*, 44–56.

Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 136-151.

Twenge, J. M., Haidt, J., Lozano, J., & Cummins, K. M. (2022). Specification curve analysis shows that social media use is linked to poor mental health, especially among girls. *Acta Psychologica*, *224*.

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *Plos One*, *18*, e0291668.

U.S. Surgeon General (2023). Social media and youth mental health. Available at: https://www.hhs.gov/surgeongeneral/priorities/youth-mental-health/social-media/index.html

Valkenburg, P. M., Meier, A., & Beyens, I. (2022). Social media use and its impact on adolescent mental health: An umbrella review of the evidence. *Current Opinion in Psychology*, *44*, 58-68.

Valtorta, N. K., Kanaan, M., Gilbody, S., Ronzi, S., & Hanratty, B. (2016). Loneliness and social isolation as risk factors for coronary heart disease and stroke: Systematic review and meta-analysis of longitudinal observational studies. *Heart*, *102*, 1009-1016.

Van der Linden, S. (2023). *Foolproof: Why Misinformation Infects our Minds and How to Build Immunity*. New York, NY: WW Norton.

van Lange, P. A. M., & Columbus, S. (2021). Vitamin S: Why is social contact, even with strangers, so important to well-being?. *Current Directions in Psychological Science*, *30*, 267-273.

von Schenk, A., Klockmann, V., & Köbis, N. (2023). Social Preferences Toward Humans and Machines: A Systematic Experiment on the Role of Machine Payoffs. *Perspectives on Psychological Science*, in press.

Wang, F., Gao, Y., Han, Z., et al. (2023). A systematic review and meta-analysis of 90 cohort studies of social isolation, loneliness, and mortality. *Nature Human Behaviour*, *7*, 1307-1319.

White House (2023). FACT SHEET: President Biden issues executive order on sage, secure, and trustworthy artificial intelligence. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/#:~:text=October%2030%2C%202023%20FACT%20SHEET%3A,Releases%20Today%2C%20President%20Biden%20is

Wietzke, F.B., & McLeod, C. (2013). Jobs, Wellbeing, and Social Cohesion: Evidence from Value and Perception Surveys. *World Bank Policy Research Working Papers,* 6447.

Williams, D. R. (2005). The health of U.S. racial and ethnic populations. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 60*, 53–62.

Wu, T. (2016). *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. PRH Knopf, New York.

Yang, K.C., & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. *Available at https://arxiv.org/abs/2307.16336*

Yang, Y., Davis, T., & Hindman, M. (2023). Visual misinformation on Facebook. *Journal of Communication, 73*, 316-328.

Yin, J., Ngiam, K. Y., & Teo, H. H. (2021). Role of artificial intelligence applications in real-life clinical practice: systematic review. *Journal of Medical Internet Research, 23*, e25759.

Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances, 8*, eabk2607.

Zuboff, S. (2023). The age of surveillance capitalism. In *Social Theory Re-Wired* (pp. 203-213). Routledge.

Zuiderwijk, A., Chen, Y. C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*, 101577.