European Commission

Research and Innovation

The impact of Artificial Intelligence on our societies

Independent Expert Report

The impact of Artificial Intelligence on our societies

European Commission
Directorate-General for Research and Innovation
Directorate D — People: Health and Society
Unit D3 — Health and Societal Transitions
Contact Michalis Moschovakos
Email michail.moschovakos@*ec.europa.eu*
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Manuscript completed in April 2025
1st edition

PDF ISBN 978-92-68-36759-9 doi:10.2777/2377840 KI-01-26-022-EN-N

EUROPEAN COMMISSION

The impact of the Artificial Intelligence on our societies

edited by Ewa Luger and Marianthe Stavridou

2026

Directorate-General for Research and Innovation
Horizon Europe

EN

Table of Contents

1. Introduction

1.1. The issue at stake and the objective of this report

Artificial Intelligence (AI) is rapidly transitioning from a niche area of technological innovation to a pervasive force that is reshaping economies, institutions, and daily life. Its applications are now embedded in core functions across sectors—from finance, education, and healthcare to communication, mobility, and public administration. As AI technologies become more sophisticated, particularly with the development and deployment of large language models (LLMs), they are unlocking unprecedented capabilities in data-processing, pattern recognition, and automated decision-making. These advances promise to enhance productivity, accelerate scientific discovery, and optimize service delivery. At the same time, they raise fundamental questions about fairness, accountability, and societal impact.

This report emerges at a critical moment in the evolution of AI policy and research. The accelerating pace of AI integration has outstripped the development of comprehensive frameworks for understanding and governing its impacts. Whilst AI offers substantial opportunities for innovation and efficiency, it also presents disruptive challenges—especially in terms of employment structures, the distribution of economic value, and shifts in institutional authority. Automation and algorithmic systems are not only transforming how work is performed but are also influencing which roles are valued, who holds decision-making power, and how benefits are shared across society. Moreover, AI's influence on democratic governance, surveillance capacities, and the spread of misinformation poses significant concerns for public trust and the resilience of civic institutions.

The objective of this report is twofold. First, it provides a critical assessment of the current research landscape surrounding AI, identifying both strengths and gaps in existing knowledge. This includes an exploration of under-researched areas that warrant greater attention from academic communities, industry actors, civil society, and policy makers. Second, the report evaluates the societal and economic consequences of AI integration, with particular focus on three key dimensions: labour markets and employment, inequality and wealth distribution, and the ethical and institutional frameworks required to govern AI responsibly.

Healthcare is used throughout the report as a thematic lens to illustrate both the promises and perils of AI deployment. As one of the most data-intensive and ethically sensitive sectors, healthcare provides a revealing case study of how AI can dramatically improve diagnostic accuracy, personalize treatments, and increase system efficiency. However, these benefits come with significant risks—such as eroding patient privacy, undermining clinician autonomy, and creating dependencies on opaque algorithmic systems. Balancing innovation with human-centric care, safety, and trust is a recurring theme in the analysis.

Grounded in interdisciplinary expert review, this report puts forward a series of strategic research recommendations aimed at informing Horizon Europe's future agenda. In addition to outlining areas for further academic investigation, it presents a coherent set of policy proposals designed to ensure that AI development aligns with Europe's values of inclusivity, transparency, and democratic accountability. These recommendations seek to bridge current knowledge gaps, enhance societal preparedness, and promote responsible AI innovation that benefits all segments of society.

2. State of play of research on Artificial Intelligence on our Societies

2.1 Mapping research on Artificial Intelligence

Global Trends in Artificial Intelligence Research and Development: The 2024 AI Index Report by Stanford University offers a comprehensive analysis of the evolving landscape of artificial intelligence (AI), highlighting major trends and developments shaping the field. A significant trend observed is the dominant role of industry in driving frontier AI innovation. In 2023, industry spearheaded the development of 51 notable machine learning models, while academia contributed only 15. However, collaboration between academia and industry has reached unprecedented levels, with 21 joint AI models developed in the same year. The growth of foundation models has also been exponential, with 149 models released in 2023, 65.7% of which were open-source, emphasising a global shift toward transparency and shared knowledge.

Geographical Leadership in AI Research and Patents: The United States continues to maintain its leadership position in AI research and development. In 2023, U.S.-based institutions were responsible for 61 notable AI models, significantly outpacing the European Union's 21 and China's 15. Patent activity further highlights the sector's vitality, with a 62.7% increase in global AI patents between 2021 and 2022, marking a staggering 31-fold increase since 2010. China leads in patent origins, accounting for 61.1% of global AI patents. Open-source AI is also flourishing, with over 1.8 million AI projects hosted on GitHub by 2023, indicating robust community engagement. Academic research has expanded considerably, with AI-related publications nearly tripling since 2010, growing from approximately 88,000 in 2010 to around 240,000 in 2022. The United States, China, and India collectively account for over 60% of global AI research output from 2010 to 2022. Strong collaboration networks, particularly between the U.S. and China, have been a defining feature of this research, with key themes including machine learning, deep learning, and natural language processing. Additionally, emerging fields such as ethical AI, AI applications in climate change, and healthcare are gaining traction (Mardiani & Iswahyudi, 2023).

A Paradigm Shift in AI Innovation: DeepSeek, an emerging AI powerhouse, has made remarkable advancements in large language models and multimodal Artificial Intelligence. In 2023, the company introduced cutting-edge models that demonstrated superior reasoning, natural language understanding, and multi-turn dialogue capabilities (DeepSeek, Xiao Bi *et.al.*, 2024). DeepSeek's commitment to open-source AI and collaborative research has reinforced global AI democratisation efforts, challenging established leaders in the field. The latest developments of January 2025 indicate that DeepSeek has surpassed Western AI leaders not only in performance but also in efficiency. In Europe, Mistral[1], a French start-up emerged, as one of the future players in AI field. The company is focused on developing advanced, open-source language models that aim to provide competitive performance and transparency in a landscape often dominated by a few large, proprietary systems.

AI Developments in the European Union: The European Commission AI Landscape Dashboard (AI Watch[2]) reveals an expanding AI startup ecosystem, with increasing investments and funding rounds fuelling AI-driven businesses, especially in healthcare, robotics, and autonomous systems. Between 2013 and 2023, AI research publications in the European Union rose by 300%, from approximately 2,500 in 2013 to over 10,000 in 2023 (Frankowska & Pawlik, 2022). Germany, France, and the Netherlands lead AI research in the region, contributing over half of the EU's AI output. International collaboration remains a hallmark of EU research, with nearly 20% of publications co-authored with non-EU researchers. Thematic priorities in European AI research include machine learning, deep learning, and AI ethics, with healthcare and industrial automation emerging as top application areas. The Horizon 2020 program is fostering AI research and

[1] https://mistral.ai/
https://ai-watch.ec.europa.eu/tools/ai-landscape-dashboard_en

collaboration, contributing over €1.5 billion to AI projects (Frankowska & Pawlik, 2022; AI Watch). Despite efforts, the European Union lags behind global leaders, and data suggest a discrepancy in research between northern and southern EU countries (Frankowska & Pawlik, 2022; Eurostat, 2024). Recent data show that in Greece, the University of Patras (1857% rise, ranked first) and Harokopio University (507% rise, ranked eighth) are among the ten fastest-rising AI grant winners in Horizon Europe (Science|Business, 2024).

Emerging AI Leaders: Switzerland has established itself as a leader in AI ethics and research, with world-class institutions such as ETH Zurich, EPFL, and the Universities of Zurich and Geneva making significant contributions to AI development. These institutions excel in fields like healthcare AI, robotics, deep learning, and AI ethics. National initiatives, including the Swiss National AI Institute, and cutting-edge infrastructure like the Alps supercomputer further bolster Switzerland's research impact. Notably, nearly 40% of Swiss AI[3] (2025) publications result from international collaborations, underscoring the country's pivotal role in global AI partnerships. Switzerland's commitment to ethical AI ensures that innovations align with human values, fostering trust and responsibility in AI development[4]. With annual investments ranging between CHF 1 billion and CHF 2 billion, Switzerland successfully bridges the gap between academia and industry. Norway is also emerging as a key player in AI research, driven by leading universities such as NTNU and the University of Oslo. Government support and a thriving tech ecosystem, including initiatives like NORA[5], have positioned Norway at the forefront of AI development. The country focuses on fields such as natural language processing, energy optimisation, and AI for climate sustainability. With a strong emphasis on ethical AI, data privacy, and transparency, Norway is establishing itself as a hub for sustainable and impactful AI innovation.

The Competitive Global AI Landscape: China and the United States continue to dominate AI research and development, fiercely competing with vast investments in R&D. Industry leads in AI innovation, leaving academia struggling to keep pace due to the high costs of training AI models and geographical disparities in research output. However, the emergence of cost-effective, open-access AI models such as DeepSeek in China challenges the expensive, closed-source models developed in the U.S. Meanwhile, the European Union remains at a disadvantage despite efforts in Horizon 2020 funding, as France announced Recently France announced 109 billion Euro investments in the sector.[6] Non EU countries in contrast, like Switzerland and Norway are using substantial funding to set high standards for an ethical and collaborative AI.

The evolving AI landscape underscores the importance of balancing innovation, accessibility, and ethical considerations as nations and organisations strive to lead in the race for AI supremacy. Overall, these insights illustrate a rapidly evolving global AI ecosystem, characterized by increased industry-academia collaboration, growing investment in AI startups, and a shifting focus on ethical, transparent, and impactful AI solutions. The integration of funding data, startup activity, and sector-specific innovation from the European Commission's dashboard provides a comprehensive understanding of how AI is shaping industries, driving economic growth, and addressing global challenges (AI Watch).

2.2 Identifying gaps in existing research

The *2024 AI Index Report* and related studies highlight significant progress in artificial intelligence research, showcasing trends in innovation, collaboration, and application. However, these developments also reveal critical research gaps that must be addressed to ensure AI's potential is fully realized in an inclusive, equitable, ethical, and impactful manner.

[3] https://www.swiss-ai.org/
[4] https://www.swiss-ai.org/,
[5] https://www.nora.ai/
[6] https://www.lemonde.fr/en/economy/article/2025/02/10/ai-with-the-announcement-of-a-109-billion-investment-macron-intends-to-take-on-the-us_6737985_19.html

One of the most prominent gaps is the disparity between industry and academia in driving frontier AI research. In 2023, industry was responsible for creating 51 notable machine learning models, compared to only 15 from academia. Whilst collaboration has increased—evidenced by 21 joint models—academic institutions face persistent limitations due to insufficient funding and access to cutting-edge resources and academics working with industry are tied into heavy NDAs that put impediments like delays to publications, restrictions on what can be discussed or Intellectual property issues. Industry dominates AI in multiple ways: accounting for 67% of AI funding in the EU, receiving large government subsidies in the US, capturing 70% of AI-specialized PhD graduates in the US, and producing 96% of large AI models—up from just 11% in 2010 (Ahmed, 2023). Bridging this gap will require equipping academia with the infrastructure and support needed to pursue independent, impactful research.

Another pressing issue lies in the financial cost of training advanced AI models. OpenAI's GPT-4 and Google's Gemini Ultra required investments of $78 million and $191 million, respectively, which presents a major barrier for smaller organisations and underfunded regions. Research into cost-efficient training, hardware optimisation, and sustainable computing is falling behind. In contrast, DeepSeek has developed high-performing models at significantly lower costs (~$5.5 million), showing that accessible, resource-conscious innovation is possible[7] (Conroy & Mallapaty, 2025).

Geographic imbalances also persist. The United States produced 61 significant AI models in 2023, outpacing the combined output of the EU, UK, and China. Even within the EU, research capacity remains uneven. Whilst AI ethics is increasingly emphasized, efforts to develop harmonized global governance frameworks addressing bias, transparency, and accountability are still in their infancy.

Meanwhile, open-source AI initiatives are gaining traction—65.7% of foundation models released in 2023 were open-source—but concerns about misuse, security vulnerabilities, and lack of quality standards remain. There is an urgent need for research into protocols, evaluation frameworks, and responsible use policies.

There are also thematic blind spots. Whilst machine learning, deep learning, and NLP dominate the field, areas such as AI for climate change, sustainability, and social good—especially for marginalized communities—are still underdeveloped. Interdisciplinary collaboration could accelerate progress here.

Global inclusivity remains a challenge. Although strong ties exist between dominant players like the US and China, developing nations often struggle to participate in international AI research efforts. Closing this gap requires equitable access to resources, funding, and expertise.

Finally, the AI startup ecosystem is booming—especially in healthcare, fintech, and autonomous systems—but scaling these innovations to meet real societal challenges is difficult. There is limited research into long-term impacts in sectors like education and agriculture, and broader implications such as job displacement, changing labour markets, and economic inequality are still not well understood. Research into predictive frameworks and mitigation strategies is crucial to ensure responsible integration of AI across society.

2.3 Research Recommendations

Private industry is likely to continue driving frontier AI research through massive investment in ever-larger datasets and models. To ensure a more balanced, ethical, and inclusive research landscape, the following recommendations focus on supporting public research, reducing regional disparities, and reinforcing trust in AI systems.

[7] https://spectrum.ieee.org/deepseek

Empowering Public Institutions in AI Development: Publicly funded institutions must play a larger role in AI development. Recent examples —such as DeepSeek's efficient models— prove that high-quality research can be achieved at relatively low cost. Academic institutions, public-private partnerships, and cooperative research models should be leveraged to support long-term projects focused on ethical and risk-mitigating AI — an area often overlooked by the private sector due to a lack of immediate profit potential.

Building a Pan-EU Open-Source AI Research Platform: A key strategy is to establish a collaborative EU-based AI research platform to support ethical research and under-resourced institutions. This would enable the formation of diverse consortia, improve access to datasets, and foster more equitable distribution of capabilities across the region. Additionally, making AI technologies developed through public initiatives open-source would help prevent monopolisation by private entities and increase transparency.

Support for cooperative AI endeavours: AI cooperatives are emerging as a robust model for AI development and adoption. However, there is a critical need for research to understand best practice and seed funding to establish the tools and conditions under which such communities might take advantage of AI.

Moving towards sovereign/EU efforts: Efforts should be made towards reducing domestic reliance on US big tech, whilst promoting international collaboration and multilateral governance. This will set a direction of travel towards a more progressive and sustainable global innovation climate. Research should identify robust routes to software and AI sovereignty, sufficient that Europe's innovation culture and potential is insulated from shocks in global politics.

Integrating Humanities for Ethical AI Insights: Ethical AI research should also include scholars from the humanities to provide deeper insight into the societal impacts of emerging technologies. This interdisciplinary inclusion can help identify and address ethical blind spots that purely technical teams may overlook.

Supporting development of minority and low-resource AI: There is currently a global underinvestment in, and politicisation of EDI coupled with an absence of languages considered low-resource. When it comes to AI as a sociotechnical tool, further investment is required to develop coproduced AI tools and models that reflect, rather than undermine or eradicate, social diversity in all its forms.

Prioritising Transparency to Foster Public Trust: Transparency in AI decision-making must become a research priority. As the internet populates with both idealized and synthesized content, assessing authenticity and veracity is falling beyond human ability. Research is required to support transparency and literacy to support identification of generative AI in use. It is critical for improving public confidence, especially in high-stakes fields like healthcare, education, and policing and to arrest democratic backsliding. Research into public perceptions and trust-building strategies will be essential to ensure AI systems are not only accurate but also perceived as legitimate and fair.

AI literacy as essential for contemporary democracies: Low levels of AI literacy are enabling democratic backsliding. This is no simple matter as use of AI requires varying levels of user comprehension. Further research is required to map existing efforts and understand the requirements and limits of AI literacy. This should include extend beyond skills to include support for the development of robust user mental models, new rules to support meaningful user experience design, understanding of available routes to citizen recourse and repair, and a return to and refresh of media literacy in the context of AI.

Support for technical and social solutions to provenance: The slackening of safeguards has added to an environment within which identifying whether platform content is factual and authentic becomes highly challenging. Further research is required to support both human and machine identification of content provenance and veracity.

Mobilising Public Sector Leadership in AI R&D Funding: Addressing the current imbalance in R&D funding requires stronger public sector leadership. Governments and academic bodies must proactively support initiatives that prioritize public good over short-term profit. This includes developing AI models that are safe, inclusive, and geared toward solving societal problems rather than solely commercial ones.

Mitigating AI's Social and Economic Disruptions: Urgent attention is needed to understand and mitigate AI's social and economic disruptions. This includes research on job displacement, changing labour dynamics, and the design of retraining and reskilling programs. Without robust, forward-looking policies and academic guidance, AI could deepen inequality and trigger instability. Researchers and policymakers must collaborate on frameworks that protect vulnerable workers and promote lifelong learning in an AI-driven economy. Further research is required to understand that long term impacts of AI in social, economic and political life.

Empower informed trade-offs: Whilst any tangible benefits of AI are still mostly theoretical, the environmental costs are real and demonstrable, suggesting that further research is needed both to demonstrate the true domain-specific value of AI, but also the environmental impact so that informed trade-offs can be made by those seeking to use the technology. Organisations will seek to understand how use of AI will balance against their commitments to sustainable development.

3. Discussion: Artificial Intelligence applications and Implications

This section provides a definition of AI and examines its application across various sectors of society, the economy, and anticipated impacts. It unpacks the ethical implications of AI integration regarding misinformation, AI decision making, responsibility and governance and explores AI's role in enhancing productivity and efficiency, alongside its potential to contribute to economic disparities. We consider how AI might provide greater benefits to certain groups, including high-income earners, capital and business owners, with implications for gender, wealth, and ethnic inequalities. Special attention is given to the arts and services sectors, analysing how AI could shape human creativity. Additionally, the section addresses the potential for economic polarisation among EU Member States. Strategies will be outlined to promote equitable AI-driven economic growth, including policies for reskilling workers in order to foster a just transition.

3.1 The Ethical Implications of AI

As with any socially embedded technology, artificial intelligence raises ethical concerns and calls for normative responses. Questions of how and when we should make use of AI, the limitations and moral implications of such systems, and how we best govern, regulate, and control AI innovation, and its effects, have been especially prominent over recent years. Concerns over alignment–how we align technology to human values–has long been a core question and is particularly salient when discussing AI as illustrated by the near unmappable rise of ethical frameworks and principles across all sectors. The drive towards identifying and agreeing the values that should guide AI, drawn initially from bioethics, highlighted beneficence, non-maleficence, respect for autonomy, and justice as the four leading principles. These have since broadened to reflect the challenges posed by AI systems to include accountability and liability, dignity and human rights, the rights of children/adolescents, diversity and inclusion, democratic values, education/literacy, human-centeredness, the protection of intellectual property, labour rights, cooperation/fair competition/open source, privacy, safety/security/reliability/trustworthiness, transparency/explainability/auditability, and truthfulness (Corrêa *et.al.*, 2023). Before we explore some of the current ethical issues surrounding AI, it is helpful to establish a common understanding of the term.

3.1.1 Artificial Intelligence

Artificial Intelligence (AI) is a term that describes a class of tools, methods and approaches designed to perform tasks typically requiring natural intelligence; for example, functions such as speech, learning, reasoning, prediction and translation. Since Alan Turing, the broad goal of AI researchers has been not only to replicate intelligence, but to create synthetic systems that could convincingly perform tasks in ways indistinguishable from a human. In its nascence, the field of AI was concerned with enhanced problem solving and the replication of skills that humans had already mastered (e.g., symbolic AI). Such systems, still widely used, require explicit programming and are human-readable, such as expert systems designed to simulate the expertise or behaviour of a human or organisation within a given domain. These systems are highly constrained, draw from an existing knowledge base to infer likely outcomes, and are regularly used in domains such as financial services, transportation and healthcare. Since this time, the field of AI has undergone several evolutions, the most notable being the development of machine learning and more recently Generative AI (GenAI), with each evolution having moved the field further away from human oversight through a growing emphasis upon system agency.

3.1.2 Machine learning and generative AI

Machine learning (ML), a term with which most people are familiar, adapts to a range of inputs, both human and data, allowing for ever more sophisticated insights, and in the case of deep learning even training itself to detect patterns. To date, it is this type of predictive model we would most commonly see embedded within products available to us on the market. Deep learning, an extension of ML, is based on neural networks–approaches that seek to replicate the way neurons operate in the human brain–enabling them to 'learn' and make increasingly complex predictions even with new datasets. The term 'foundation models' describes large deep-learning neural networks that, having been trained on a broad range of data, can be more easily adapted to perform a range of tasks (Thieme *et al,* 2023). To operate, such powerful models require both huge training datasets and a high level of computer power, and this is particularly true of the latest evolution in AI, GenAI which has seen a rapid expansion since 2019 (Goktas, 2024).

GenAI refers to a branch of deep learning that uses a very specific and complex type of neural network, making use of hundreds of billions of neurons. As the name suggests, GenAI is a type of AI that *generates* new content (e.g., images, text, audio, video, code) and uses that knowledge to sythesize new content that fits the pattern of what it has previously learned. Unlike AI in the past, these new models operate at incredible speed and scale, generating content in seconds to minutes. Consider the voice assistance on your phone, an online translation tool or Google search, each of these applications makes use of GenAI in the form of Large Language Models (LLMs). The newest iterations of GenAI we are seeing, such as GPT4 (Generative Pre-Trained Transformer), are built upon far more sophisticated models that require user prompts (sets of instructions written as we would speak, in natural language), allowing us far more complex and sophisticated outputs. Whilst prior models were embedded in existing products, models like GPT4 are available as tools directly to the user, free at the point of use and have seen a staggering uptake. For example, launching in November 2022 ChatGPT secured 1 million users within its first 5 days and, as of November 2024 claims to have 100 million weekly active users[8].

[8]Mortensen, O. (2025) *How Many Users Does ChatGPT Have? Statistics & Facts (2025)* https://seo.ai/blog/how-many-users-does-chatgpt-have#:~:text=ChatGPT%20experienced%20a%20meteoric%20rise,active%20users%20as%20of%20November (last accessed 06.01.25)

3.1.3 Black box AI as a sociotechnical, political, and environmental concern

AI systems do not operate in a vacuum; they are shaped by humans at every stage of the workflow. AI tools are trained on human-generated content that apply logic and instructions built into systems by humans, then also evaluated, refined and used by humans. They are trained on human data that, without intervention, brings with it and amplifies all our historical biases such as outdated beliefs about race and gender. The most recent evolution of AI has seen systems designed with the capacity to autonomously learn and adapt at speed and scale. As such, the underpinning logic is not readily inspectable, raising questions around transparency of decision-making and problematising accountability. This has framed AI systems as back boxes, unaccountable and uninspectable. Whilst it is tempting to prioritise technical solutions, AI systems are actually complex sociotechnical assemblages involving both social and technical practices, culture, politics, institutions and infrastructures (Crawford, 2021). AI systems do not operate in isolation but are always embedded within complex socio-political and economic contexts and organisations, requiring broad multidisciplinary solutions to ensure responsible and ethical innovation and use (Oduro & Kneese, 2024). According to David Runciman (2023), the notion of a black box system is not unique within human history. Black boxes are in fact all around us in the form of corporations and governments and, in this regard, AI is not special. We must apply the same checks and balances that we would to any complex sociotechnical system (Runciman, 2023). The narrative of AI uniqueness, coupled with the trend towards uninspectability has, arguably, served its designers well as a form of marketing (Milne, 2020). By relying on magical narratives, anthropomorphism, hype and stimulating fear of missing out (FOMO) (Kurtzig, 2025) they have forced a climate within which AI innovation is rarely directly challenged, allowing for maximum profit with minimum responsibility.

> [AI] is built to maximize the extraction of wealth and profit – from both humans and the natural world–a reality that has brought us to what we might think of it as capitalism's techno-necro stage. In that reality of hyper-concentrated power and wealth, AI–far from living up to all those utopian hallucinations–is much more likely to become a fearsome tool of further dispossession and despoliation[9]. (Naomi Klein, 2023).

Whilst AI is a relatively recent focus, the notion that technology is inherently political is not new. Lawrence Lessig (2006) famously asserted that decisions made by those who design and control computer systems ("code") have profound impacts on behaviour and society in ways akin to law. He identified it as one of four key forces that shape human behaviour of which technology; Law, norms, markets and architecture (code). As the latest iteration of *'code'*, AI is not only shaping our behaviour but also our geopolitics as nations race to dominate the field.

As a consequence, the environmental cost of large models often fails to attract mainstream attention or capture public imagination. AI data centres produce electronic waste, use significant amounts of energy, consume huge amounts of water, and rely on minerals and rare elements that are not always sourced ethically (United Nations Environment Programme, 2024). Such costs should not be overlooked particularly, as noted by Satya Nadella CEO of Microsoft, as AI has yet to demonstrate any practical utility beyond constrained challenges (Maxwell, 2025). So, whilst the benefits of AI are still mostly theoretical, the environmental costs are real and demonstrable, suggesting that nations should pursue a proportionate and targeted approach to AI adoption, rather than chasing innovation at any cost.

3.2 AI-generated content, accuracy and bad actors

Whilst errors in translation, search and autocorrect are usually contextually constrained and low risk, new GenAI tools are being used to generate huge swathes of convincing content for ever-emerging contexts. The core concern here is that we currently have no robust way of identifying

[9] https://www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein

whether any given content is AI generated, and no fast way of ensuring accuracy or veracity. Whilst content accuracy is, of course, a core goal of system designers, there are no assurances. Levels of model accuracy tend to be achieved through a series of human interventions throughout the AI lifecycle. This begins with judgements about the scale and quality of the training dataset and subsequent fine-tuning, through to more curated contextually relevant datasets. Reinforcement Learning from Human Feedback (RLHF) is a method that is used to align the trained model to a desired behaviour, context or value (e.g., reduction of bias, politeness) (Askell, *et al.*, 2021) a practice known as alignment. Beyond this, error reports from users and integration of new data enable system updates, and data annotation work, often annexed within large centres in global majority countries (Chandhiramowuli *et al.*, 2024), supports the development of more contextually accurate performance. Beyond this, models will be tested and evaluated in various ways, such as through red-teaming, and further lessons will also be learned from deployment in the field, and through filtering processes such as content moderation. It should be noted that the displacement, to the global south, of processes such as data annotation and content moderation itself poses a host of ethical issues. Most recently, Meta has faced a series of lawsuits in Africa due to the psychological impacts of exposure to content that includes child sexual abuse and extreme violence (Hall and Wilmot, 2025). Outsourcing harms to the global south, both social and environmental, is currently one of the tolerated operating costs of AI and should be more explicitly included when weighing the moral costs of AI adoption.

Still, even when such methods are employed, systems can still exhibit unexpected and undesirable behaviours at the point of use, particularly when exposed to bad actors. For example, in 2016 Microsoft released Tay, and AI chatbot that learned from external input. Based upon the success of Xiaoice, a similar product already deployed in China, Tay was released on Twitter and within 16 hours of being exposed to bad actors the model was openly expressing racists, anti-Semitic, and anti-feminist content making it not only a classic cautionary tale, but also an example of the difference that cultural context makes at the point of deployment. Whilst Microsoft claimed the model had been stress-tested (Lee, 2016), online attacks had surfaced unexpected vulnerabilities resulting in the generation of offensive content.

3.2.1 AI hallucinations and output veracity

Moral concerns over GenAI are not solely a product of bad actors. The models themselves can be unreliable, falling short of the standards of veracity and accuracy expected of human actors. Examples of OpenAI's Whisper transcription tool, whilst claiming near-human levels of accuracy, has been found to invent text and sentences that were never spoken[10]. This type of error is so frequently occurring in GenAI powered systems that it has become known as a 'hallucination'[11]. Whilst companies like OpenAI have been clear that particular GenAI products should not be used in high-risk industries,[12] such recommendations are currently buried within content policies rather than flagged upfront. To offer some examples, a complaint was made to the Norwegian Data Protection agency that ChatGPT had falsely stated that a man had murdered his two sons and was subsequently jailed for 21 years. This was a factual inaccuracy, and the victim has called for OpenAI to be fined, citing concerns over people's increasing belief of fake news and the possible reputational damage. In January 2025 Apple intelligence news was also suspended for making repeated errors within both headlines and news summaries (Rahman-Jones, 2025), and a recent study has shown that AI search engines cite incorrect news stories at around 60% of the time —

[10] https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14

"AI hallucination is a phenomenon wherein a large language model (LLM)—often a generative AI chatbot or computer vision tool—perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate." (IBM) https://www.ibm.com/think/topics/ai-hallucinations

[12] https://openai.com/policies/usage-policies/

"The study highlighted a common trend among these AI models: rather than declining to respond when they lacked reliable information, the models frequently provided plausible-sounding but incorrect or speculative answers—known technically as confabulations. The researchers emphasized that this behaviour was consistent across all tested models, not limited to just one tool" (Edwards, 2025). This research showed that such tools also tended to direct users to syndicated content versions over the actual source.

It should be noted that the effects of poor veracity have also progressed beyond the social, to economic impact. In 2023 Google's AI chatbot, Bard, hallucinated that the James Webb Space Telescope had been the first to capture images of a planet outside the Earth's solar system. A claim they included in their promotional video, resulting in a loss of £82bn for their parent company Alphabet as investors fled.[13] The rise of this kind of slop, the term used to describe AI-generated content (text, images, media), across the Internet and social media is also likely to desensitise citizens to the authenticity of all content–an outcome further reinforced by trends embedded in the design of social media platforms.

Researchers at Stanford University found that slop was upranked (actively promoted by the engagement algorithm) by Facebook in users' feed, despite those users being unaware that the content was synthetic (Diresta & Goldstein, 2024). Whilst GenAI content is still technically discernible in many cases, its proliferation is likely to have a desensitising effect, particularly as human-generated content on social media is equally adjusted to align with more idealised user-preferred aesthetics (Ozimek *et al.*, 2023) such as facial symmetry, skin lightening and slimmer body size. Visual preferences that are also reflected in AI-generated content. As the internet populates with both idealised and synthesized content, assessing authenticity and veracity will fast fall beyond human ability.

3.3 Misinformation and disinformation

The democratisation[14] of GenAI tools, such as Google Bard and ChatGPT, has already had a substantial impact upon the written word. Whilst such tools can be used as summative or interpretive instruments, making complex text more accessible, they are also beginning to have profound impact upon fields such as education, scientific publishing (Bagenal *et al.*, 2024), news and journalism. To assess the scale of the problem, Haider *et al*., (2024) sought to determine the extent to which GPT-fabricated papers were present in existing commonly used archives, citation databases, social media platforms and repositories. They found that approximately two-thirds of the papers they retrieved had been (either entirely or partly) generated by GPT without this being disclosed by the authors. The researchers further note that 57% of those papers were about subjects of interest to policymakers, such as health, computing and the environment (ibid), raising questions as to the potential influence of such content. There are, however, some efforts to combat this trend. For example, the International Committee of Medical Journal Editors have gone so far as describe how GenAI should be used, whilst the Lancet, a highly reputable scientific journal, asks authors to be explicit about the LLM they used, including the version and prompts, in addition to specifying why and where in the manuscript it was used. (Bagenal *et al.,* 2024).

Whilst the presence of algorithmically generated academic papers existed prior to GPT (Cabanac & Labbé, 2021), the public availability of LLMs has placed such tools in the hands of anyone with an internet connected device. Recent research has shown that pollution of the data commons is also present within informal online publishing and has increased in the past couple of years (Knibbs, 2024). An analysis of content on the blogging platform Medium, by AI startup Pangram Labs, stated that 47% was likely to have been AI generated, and that one day of global news sites (summer 2024) included 7% of synthetic content (Knibbs, 2024), either wholly or in-part. A further

[13] https://www.theguardian.com/technology/2023/feb/09/google-ai-chatbot-bard-error-sends-shares-plummeting-in-battle-with-microsoft

[14] A term often used to describe the broad availability and accessibility of a technology

study by Originality AI, an AI detection startup, offered similar estimates. AI detection tools, whilst helpful, are themselves also subject to false positives and negatives. The speed of AI innovation also means that as soon as a detection tool is available, the market responds by developing new tools to trick them. Paraphrasing tools, designed to more closely mimic human-generated output (such as Paraphraser[15] and Quillbot)[16] are now commonly used, making detection far harder, to the extent that "there is currently no technology that can reliably identify the use of generative AI in articles and so publishers rely on authors transparently declaring its use" (Bagenal *et al.*, 2024 p2142).

3.3.1 Transparency

Transparency is a core tenet of ethical and responsible AI design. It describes both a goal, and the means by which one might understand the underpinning logic of an AI system. Effective transparency must take into consideration the context, expertise of the user, and level and means of logic-exposure necessary to the task (Simkute *et al.*, 2021). Whilst factors such as algorithms being proprietary information, or the complexity of the systems, have slowed down efforts to achieve this in many cases, the newly agentic nature of deep learning has made transparency near impossible. In practice, such models have become obscure even to subject experts or designers of those systems. Transparency, therefore, remains an unsolved problem.

However, achieving effective transparency is key to developing mechanisms to combat misinformation (accidentally incorrect information) and disinformation (purposely incorrect information) as well as in building user trust (Wachter & Brynjolfsson, 2024). Misinformation and hallucinations problematize the use of AI in sensitive and high-risk domains such as healthcare, national security and policing, in addition to lower risk but still critical domains such as science (ibid). As deep fakes–convincing content that is wholly or partially generated by deep learning- and mis/disinformation have become increasingly prolific (Shoaib *et al.*, 2023), a lack of transparency is likely to further compound the pollution of our global information ecosystem.

3.4 Manipulation, democracy and epistemic rights

Synthetic content has already begun to affect public opinion globally (Swenson & Chan, 2024) to the extent that the state of California recently put in place laws to make it illegal to both create and publish election related deepfakes 120 days before election day, and 60 days after it, in order to "safeguard the integrity" (Nguyen, 2024) of the electoral process. Whilst such content may still be detectable in some cases, AI innovation moves at pace, and it is notoriously difficult to track the originators and hold them to account (Swenson & Chan, 2024). AI generated content is having similarly profound consequences for media and journalism, with the potential to erode public trust (Ciftci, Yuksek & Demir, 2023). As online platforms such as Twitter/X and Facebook have become an increasingly important means by which citizens access news, the concurrent rise of AI generated content has created a fertile climate for misinformation. The recent slackening of safeguards through the abandonment of fact checking by Meta (Caro, 2025), and content policing by X, have added to an environment within which identifying whether platform content is factual and authentic becomes highly challenging.

3.4.1 US dominance and European tech sovereignty

In the United States, the January 2025 rollback of protections put in place by the Biden administration, such as the 'Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'[17], opened the door to a global spread of AI that falls below expected ethical standards.

[15] https://www.paraphraser.io

[16] https://quillbot.com/paraphrasing-tool

[17] https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

It is worth noting that all of the big five tech companies are based in the United States and becoming ever more ideological and politically influential. This move is in contrast to the ongoing establishment of guardrails within Europe, as nations implement the EU AI Act - most recently Spain has approved a bill that mandates substantial fines for unlabeled AI content, such a deepfakes, requiring that such content be distinguishable from the first interaction. The bill also prohibits the use of biometric data being used as training data[18]. In a move to both weaken the dominance of US software and enhance their digital sovereignty, Germany and France are now promoting a new open-source web-based collaborative tool called Docs, which is designed to offer a viable alternative to systems such as Google docs and is available to download[19]. "Docs is a joint development effort between France's Interministerial Directorate for Digital Affairs (DINUM) and Germany's Center for Digital Sovereignty of Public Administration (ZenDiS). Both agencies have the goal of funding and organising digital projects that improve digital sovereignty, and Docs is built primarily as a tool for local agencies and companies" (Davenport, 2025). Similarly, in the same month, the Dutch parliament approved motions to reduce dependence on US software companies, which also comes with longer term plans to develop a sovereign cloud services platform (Sterling, 2025). These moves are supported by the European tech industry, which also recently called for European commitment to the development of sovereign digital infrastructure, thus reducing reliance on US big tech (Lomas, 2025), aligning with the advice to reduce economic dependencies, as set out in the Draghi report and EU compass.[20] A recent example being GPT@EC, a general-purpose corporate tool, developed by the Directorate-General for Digital Services (DGIT).[21]

Europe's repositioning comes at the same time as the United States engages in intense competition to lead on AI globally. Graylin and Triolo (2025) state that, having been seen as the leader based on both scaling potential and availability of advanced compute, it is now clear that these factors no longer confer the same domestic advantage to the United States as they once did. China's most advanced models are not only less resource-intensive but present results that, on the face of it, offer near equal efficacy.

Historically, the United States had sought to neutralise this competition by narrowing China's access to resources (e.g., semiconductors), but recent Chinese AI advances (DeepSeek) have invigorated global competition. The authors further argue that, without a cooling of such tactics "the consequences could be severe—undermining global stability, stalling scientific progress, and leading both nations toward a dangerous technological brinkmanship. This is particularly salient given the importance of Taiwan and the global foundry leader TSMC in the AI stack, and the increasing tensions around the high-tech island" (Graylin & Triolo, 2025). Since this article was published, the conflict has escalated through a programme of global tariffs implemented by the Trump administration, creating yet greater tensions with China. AI security and leadership are fast becoming the defining national security issues of current times, and how Europe responds will have profound consequences for future sovereignty and security. Efforts towards reducing domestic reliance on US big tech, whilst promoting international research collaborations and multilateral governance (ibid) will set a direction of travel towards a more progressive and sustainable global innovation climate.

3.4.1 Power, democratic backsliding and personalisation

Platform businesses play an increasingly pivotal role in AI within news due to their control of the AI infrastructure upon which many news providers and consumers rely (Simon, 2022). News organisations are also more likely to work with tools developed by one of the big five tech companies (Google, Amazon, Microsoft, Meta and Apple) than to seek development of local bespoke AI systems as the overheads - access to (unbiased) training data, technical expertise,

[18]https://www.medianama.com/2025/03/223-spain-leads-eu-in-ai-regulation-unlabeled-deepfakes-could-cost-millions/

[19]https://impress-preprod.beta.numerique.gouv.fr/login/

[20]https://ec.europa.eu/commission/presscorner/detail/en/ip_25_339

[21]https://commission.europa.eu/news/commission-launches-new-general-purpose-ai-tool-gptec-2024-10-22_en

human labour, cost of development, sustainability of systems, available compute and wider infrastructure – are prohibitively high (Simon, 2022; Diakopoulos, 2019). The resulting influence of news by platform companies is now arguably close to 'media capture', a term coined by economists and used to describe the control of media by governments or vested interests.

An objective and fair news media is critical to a functioning democracy but the increasing examples of undue influence on global media services show that this is contributing to democratic backsliding (Wright *et al.*, 2024); a shift in power towards the executive, coupled with a significant erosion of democratic norms and institutions. (Russell *et al.*, 2022). Whilst AI is not directly responsible, it is a novel contributing factor, particularly as news organisations come under pressure to transform their services through its use. We are seeing growth in the use of ML and other algorithmic systems within news production, requiring those involved to sufficiently understand the logic and operation of these systems (transparency) so as to not undermine journalistic values; objectivity, accuracy and impartiality (Jones *et al.*, 2022, p.1733). This is particularly true of public service media, which is explicitly normative (ibid).

Research shows that audiences increasingly expect news to be personalised (Rezk *et al.*, 2024), and this is achieved through news recommendation systems that rely on algorithms. However, despite reporting a preference for agency in the news curation process, users are in practice unlikely to intervene in content personalisation even when offered the opportunity (ibid). As hyper-personalisation becomes increasingly common, concerns over GenAI compounding disruption through opacity of provenance and risks of inaccuracy become salient concerns for news providers (Gartry, 2024). A further challenge to our news ecosystem is disintermediation, the concern that AI-driven services, such as chat, will disrupt the traditional top-down information flow resulting in users no longer needing to go directly to publisher websites, and instead relying on AI-generated summaries. Google's move to position AI-generated summaries at the top of every search is a demonstration of what this might look like in the future, exploiting two trends; the fact that users tend only to engage with the top search outputs, rarely scrolling past the first page of results[22], and the 'News Finds Me' (NFM) perception. This latter phenomenon is the theory that people, particularly those who are younger with lower levels of formal education, increasingly tend to believe that they can stay indirectly informed about public affairs, therefore not needing to actively seek out news sources (Strauß *et al.*, 2021).

However, the effects of GenAI can, to some extent, be held at bay if organisations don't entirely cede responsibility to intelligent systems (Runciman, 2023). For example, an ethnographic study of the design of a personalised algorithm within a Danish news organisation mitigated such concerns by ensuring robust editorial control, so as not to lose the value of collective interest in the drive towards personal interest. It also showed, however, that "the otherwise strong professional ideology of journalism are renegotiated and reconfigured when encountering new actors (e.g., the algorithmic system, data scientists, data infrastructures, etc.)" (Schjøtt Hansen & Hartley, 2021, p. 936). Whilst Public Service Media organisations are in the process of aligning their policies[23], the tension between user expectation, journalistic and editorial integrity, and the availability and advances in AI make this a pressing and complex space.

[22]https://backlinko.com/google-user-behavior

[23]Dr. Kate Wright is currently mapping the use of AI within public service media globally (paper forthcoming) https://braiduk.org/responsible-ai-in-international-public-service-media

Case Study: AI Surveillance and Civil Liberties in Europe (2024–2025)

Between 2024 and 2025, several European countries faced serious controversy over the deployment of AI and surveillance tools, revealing growing risks to privacy, civil rights, and democratic freedoms.

In Hungary, authorities planned to use facial recognition AI to monitor attendees of the Budapest Gay Pride. The aim was to identify individuals and potentially issue fines. This triggered widespread criticism, as the move violated the European Union's AI Act, which bans biometric surveillance in public spaces without exceptional justification. Critics accused the government of targeting LGBTQ+ citizens and suppressing public expression[24]. In Italy, it was revealed that the government had authorized Graphite spyware to be used against members of Mediterranea Saving Humans, a refugee rescue NGO. The justification was national security, but many saw it as a clear attack on civil society and humanitarian work[25]. The Netherlands came under scrutiny for using AI to conduct predictive surveillance of travelers. Algorithms flagged individuals as high-risk without transparency, sparking concerns about profiling and bias in decision-making systems[26]. In France and Denmark, welfare fraud detection systems driven by AI were accused of disproportionately targeting vulnerable populations, including disabled individuals and single mothers. In Denmark, Amnesty International described the system as part of a broader "*automated surveillance state*".[27] Greece found itself embroiled in a political and civil rights scandal involving the illegal use of spyware, including the Predator malware, to monitor political opponents, journalists, and civil society figures. The scandal, known as "Predatorgate," brought to light the extensive surveillance operations conducted by private companies with ties to the government. This case demonstrated the use of surveillance technology as a tool for political control, undermining democratic principles and the privacy rights of citizens. The Greek government's role in the scandal continues to be a subject of debate and legal proceedings. Furthermore, the Greek Ministry of Migration was fined €175,000 by the Hellenic Data Protection Authority for deploying biometric and AI-based surveillance systems in migrant camps without adequate privacy safeguards[28]. In the context of the broader spyware debate, Denmark and Cyprus were revealed as potential customers of Israeli spyware company Paragon Solutions. Paragon's spyware was used to monitor civil society members, raising concerns about state-sanctioned surveillance practices in these countries[29]. In Poland, AI tools have been integrated into law enforcement and judicial systems, raising concerns about transparency and potential biases. The use of algorithms[30] in Polish courts to support decision-making has been noted, with experts highlighting issues related to fairness and the independence of judges. Additionally, algorithmic bias in AI systems, including those used in recruitment and credit scoring, has been identified as a concern affecting personal freedoms[31]. Poland launched an investigation into the previous government's use of Pegasus spyware, which targeted hundreds of individuals, including political opponents and journalists, raising concerns about privacy and misuse of surveillance technologies[32].

These cases collectively demonstrate the danger of deploying AI tools in ways that infringe on privacy, civil liberties, and human rights. Although often justified in the name of security or efficiency, poorly regulated AI systems can reinforce bias, target minorities, and erode public trust. As the EU prepares to enforce the AI Act, these incidents highlight the urgent need for transparency, oversight, and ethical frameworks to guide AI's use in public life.

3.4.2 Content authenticity and epistemic rights

A further impact of AI is its effect on epistemic rights. Epistemic rights are concerned with protection of "goods related to the domain of inquiry" (Della Croce, 2023, p.122), such as receiving accurate information about a medication or surgical procedure or, in line with the theme of this report, rights to information that is authentic, true and verifiable in the context of an election. In other words, we have a right to be informed as a citizen, and AI has the potential to undermine our access to accurate information and therefore also our epistemic rights.

Ensuring transparency of AI systems and their outputs is now an accepted and desirable tenet of responsible AI. However, whilst the desire is universal there are no clear or universal solutions. There are, nonetheless, a range of emerging methods of mitigation. For example, in response to content veracity and authenticity challenges posed by GenAI, the notion of synthetic content watermarking has gained traction. Google DeepMind have developed SynthID, an AI watermarking system to distinguish content developed by Gemini, a Google chatbot, from content that has been human generated (Dathathri *et al.*, 2024). In response to issues of provenance Balan *et al.,* (2023), developed a visual attribution technique that leverage NFTs (non-fungible tokens) to allow both attribution and apportioned credit, though how credit is apportioned remains an open question. The British Broadcasting Corporation (BBC) have also introduced a 'content credentials' feature, designed to indicate the provenance of an image or video and how they have verified its authenticity (Monday &. Strappelli, 2024). A self-selecting, limited user trail (1,200 people) indicated that the feature increased their trust of media to some extent. This initiative is part of the Coalition for Content Provenance Authority (C2PA)[33], an international coalition of big tech and public media organisations focused on developing technical standards for assuring the provenance of media content.

Watermarking is another technique being advanced to assess content authenticity. Akin to the analogue form, watermarking is designed to embed digital information, such as provenance and modifications, in GenAI content as a means of authentication. However, to date there have been no robust or scalable solutions. For watermarking to be effective it needs to be detectable even when images are edited, as well as sufficiently inconspicuous so at to render the image usable (Hoffman-Andrews, 2024) and there remain concerns that such methods, whilst useful, are not sufficient to curb disinformation as they are often easy to remove, must be robust enough to withstand efforts to remove them, and would only be used by good actors (Hoffman-Andrews,

[24] https://www.euronews.com/my-europe/2025/03/26/exclusive-hungarys-gay-pride-surveillance-would-breach-the-eus-ai-act-says-leading-mep?utm_source=chatgpt.com https://www.euronews.com/next/2025/03/21/from-surveillance-to-automation-how-ai-tech-is-being-used-at-european-borders?utm_source=chatgpt.com; https://www.politico.eu/article/ai-deepseek-chatgpt-openai-eu-bans-series-of-ai-practices-but-with-loopholes/?utm_source=chatgpt.com;

[25] https://www.theguardian.com/world/2025/mar/27/italian-government-approved-use-of-spyware-on-members-of-refugee-ngo-mps-told?utm_source=chatgpt.com

[26] https://www.wired.com/story/inside-the-black-box-of-predictive-travel-surveillance/?utm_source=chatgpt.com

[27] https://www.france24.com/en/tv-shows/perspective/20241114-benefit-fraud-amnesty-accuses-denmark-of-using-ai-to-build-system-of-surveillance?utm_source=chatgpt.com; https://www.wired.com/story/algorithms-policed-welfare-systems-for-years-now-theyre-under-fire-for-bias/?utm_source=chatgpt.com

[28] https://www.hrw.org/world-report/2025/country-chapters/greece?utm_source=chatgpt.com

[29] https://www.euractiv.com/section/tech/news/paragon-scandal-denmark-and-cyprus-potential-spyware-customers-alongside-italy/

[30] https://algorithmwatch.org/en/automating-society-2019/poland/

[31] https://algorithmwatch.org/en/polish-electronic-courts/

[32] https://www.theguardian.com/world/2024/apr/01/poland-launches-inquiry-into-previous-governments-spyware-use?utm_source=chatgpt.com; https://algorithmwatch.org/en/polish-electronic-courts/?utm_source=chatgpt.com

[33] https://c2pa.org

2024). It should be noted that such methods are currently experimental, small scale and do not generalise across all generative models; furthermore, they can potentially undermine privacy if all aspects of provenance are captured, highlighting the need for more robust watermarking measures.

The range of concerns and epistemic risks associated with AI in these contexts places value-led domains, such as public service media, in a position where the external pace of innovation causes considerable friction with internal values. One positive way that this is being addressed is through cross-sectoral partnership and collaboration on transferable and scalable solutions.

Case study: Cross-Sectoral Partnerships in Public Service Media

Established by Royal Charter In 1927, the British Broadcasting Corporation (BBC) is now the impartial, independent and trusted public service broadcasting organisation in the UK that "acts in the public interest, serving all audiences through the provision of impartial, high-quality and distinctive output and services, which inform, educate and entertain"[34] (BBC website).

The rise of AI in the context of news and entertainment has surfaced a series of challenges that impact public service media across its remit. As a value-led organisation, the BBC is concerned with how to use AI responsibly and in ways that align with its values. To this end, the BBC must also negotiate the delivery (and demonstration) of public value from any AI use as well as supporting public AI literacy[35]. It must also build the AI literacy of internal staff, and ensure that AI augments, rather than displaces, their jobs. The BBC has taken a proactive principle-led approach to responsible AI with three specific principles guiding work: to act in the best interests of the public, to prioritise talent and creativity, and to be open and transparent,[36] as well as developing wider guidelines for their editorial work[37]. In support of transparency and disclosure, they provide public updates related to the piloting of AI across their portfolio[38]. For example, they have published a report describing where they have successfully used AI to face-swap interviewees to protect their anonymity[39].

Partnership is another core mechanism by which the BBC addresses these issues. It has partnered with both academic institutions and technology companies, the latter evidenced through both the C2PA, and the Partnership on AI's Media Integrity strand, which is currently exploring intervention points to support the "broader quality and integrity of information online"[40]. In terms of academic partnerships, the BBC, with 13 academic partners, recently launched its Responsible Innovation Centre for Public Media Futures[41] designed to develop leadership in responsible innovation in public media, grow partnerships, and deepen research in (i) protecting and promoting public media values, (ii) shaping personalisation, (iii) promoting digital inclusion and civic participation online, and (iv) measuring public value and social impact of technology. They have also published their internal research on issues such as the impacts of AI assistants on news[42], and generated image detection[43]. A concurrent reduction in organisational income[44], coupled with the inherent risks and costs associated with use of third-party services, has meant that the BBC has had to be highly proactive and collaborative in its approach to AI development, adoption, and deployment. Trust, transparency and disclosure of AI use in BBC products and services have become dominant issues[45], as they have for many media outlets. Equally, challenges arising from the use of AI have raised concerns across related professions. For example, the impact of LLMs and GenAI on copyright and IP for creatives, the effects of disintermediation, and the rise of mis/disinformation in the context of news. These challenges require organisations to collaborate in the development of rapid and robust responses.

One such collective response is the Coalition for Content Provenance and Authenticity (C2PA). This partnership unifies two projects: the Content Authenticity Initiative (CAI), led by Adobe, that aims to deal with provenance issues through providing the context and history of digital media, and Project Origin, a Microsoft and BBC-led initiative aimed at tackling disinformation in the digital news ecosystem[46]. The overarching C2PA partnership[47] brings together key players in the field to work collaboratively towards more interoperable solutions.

3.5 AI, creative practice and intellectual property

The rising availability of GenAI has created both benefits and risks across a range of knowledge work domains, though one of the most critical areas is creative practice. In 2022 OpenAI led the

[34] https://www.bbc.com/aboutthebbc/governance/mission

launch of a series of products that captured the public imagination, placing powerful AI tools in the hands of anyone with an internet connected device. By creating a simple user-interface, free at the point of use, they lowered the barrier to entry, inviting millions to participate. Such tools have enabled people with limited creative skills to produce a range of content at low to no cost, bypassing the need for professional human creativity, or artistic skill, at the point of use.

3.5.1 The impact of AI on creative practitioners

Text to image diffusion models, such as Stable Diffusion or Midjourney, are trained on billions of images/content that are often drawn from the internet. This content was not originally made public for this purpose, and such secondary use has resulted in arguments that artistic labour being used without compensation, attribution or consent. The nature of how data is bundled for training also means that any licenses or restrictive terms associated with that data are not readily visible to those training the models (Longpre *et al.*, 2024), compounding the problem.

Whilst these tools have potential benefits for artists, but they also have the potential to harm the art workforce and infringe upon artistic and intellectual property rights. Without explicit consent from artists, Generative AI creators scrape artists' digital work to train Generative AI models and produce art-like outputs at scale. These outputs now compete with human artists in the marketplace. Whilst such models are also being used by some artists in their generative processes to create art, research shows that the majority of artists believe model creators should disclose what art is being used in AI training, that AI outputs should not belong to model creators, and that they express concerns about AI's impact on the art workforce and who ultimately profits from their art (Lovato *et al.*, 2024 p. 905).

3.5.1.2 AI, creative practice and copyright implications

According to Chesterman (2024) two intellectual property policy questions arise; whether the original creators of the content (e.g., artists) that trains the model should be compensated, and who owns the outputs of the model. The question of how AI shapes both individual and collective rights and the legal status of outputs (Jiang & Goetz, 2024) is a live issue, and of growing concern. This is particularly the case as models have been found to not only generate content that mimics an artist's distinctive style, but also that they simply regurgitate training data, which strays in the territory of copyright infringement (Lee & Cooper *et al.*, 2023). Whilst there are those with the means to contest companies' use of their creative and intellectual property, resulting in some court cases (Ren *et al.*, 2024), smaller organisations and sole practitioners can be left defenceless.

The appetite for appropriate attribution is not simply a concern of the content providers but also the professionals that use them; research conducted with user experience (UX) designers found that those who used GenAI tools in their wider design workflow felt a moral desire to attribute the outputs to the originators of the training content through some form of provenance mapping (Li,

[35]https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/consultation-ofcoms-three-year-media-literacy-strategy/responses/bbc.pdf?v=370071
[36]https://www.bbc.co.uk/mediacentre/articles/2023/generative-ai-at-the-bbc
[37]https://www.bbc.co.uk/editorialguidelines/guidance/use-of-artificial-intelligence/#3editorialissuesintheuseofai
[38]https://www.bbc.com/mediacentre/2025/articles/update-generative-ai-at-the-bbc
[39]https://partnershiponai.org/wp-content/uploads/2024/03/pai-synthetic-media-case-study-bbc.pdf
[40]https://partnershiponai.org/program/ai-media-integrity/
[41]https://www.bbc.co.uk/rd/projects/responsible-innovation-centre
https://www.bbc.co.uk/aboutthebbc/documents/bbc-research-into-ai-assistants.pdf
[43]https://www.bbc.co.uk/rd/publications/deepfake-detection-image-manipulation
[44]https://vlv.org.uk/news/bbc-real-terms-public-funding-in-2020-21-25-lower-than-it-was-in-2010-11/
[45]https://www.bbc.co.uk/rdnewslabs/news/content-credentials
[46]https://c2pa.org/"https://c2pa.org/
[47]Partners: BBC, Adobe, Amazon, Sony, Microsoft, Publicis group, OpeanAI, Intel, Google, Meta and Trupic

Cao *et al.*, 2024). As AI innovation progresses at pace the development of regulation, standards and other protective mechanisms aimed at creative practitioners' content are ever more critical. One of the more problematic features of text to image diffusion models is that they can be trained to mimic a specific artist's style by being fine-tuned on examples of their work. Given that synthetic artwork is not only being used for economic gain but has also won art and photographic awards (Metz, 2022; Grierson, 2023), this displacement of value from originator to model user is of moral concern.

In response to such challenges there have been other, more adversarial, efforts to enable artists to protect their content from being used as training data. For example, 'Glaze' is a tool developed specifically to disrupt AI mimicry through the application of a 'style cloak' which perturbs the image sufficiently to disrupt the training process, whilst still rendering it useable by the original artist (Shan *et al.*, 2023). The downside of such attempts is that this perturbation also sees a reduction of the overall quality of the output.

3.6 Responsible AI

Whilst ethics are broadly understood as tacitly agreed normative rules that guide our moral conduct, there are no agreed definitions of what we mean by Responsible AI. Recent research from BRAID, a 15.9 million Pound UK investment designed to bridge the responsibility divides between stakeholders in the AI ecosystem, define the term as 'the ambition, amongst the interconnected communities of practice that make up the AI ecosystem, to ensure AI systems are designed and deployed in ways that minimise potential risks and harms, whilst maximising opportunities for human flourishing' (Tollon & Vallor, 2025). From this perspective, Responsible AI describes a collective desire that encompasses multiple efforts.

The term is used variously to describe (a) an interdisciplinary field of academic/industry research, (b) a stated corporate governance agenda, (c) a desired type of AI product, and (d) a broad community or ecosystem of stakeholders (ibid). However, for the term to have utility it requires the ecosystem to both agree a coherent definition, and to be able to point towards tested tools and approaches that help us to identify whether or not something is designed responsibly.

To say that a system is designed responsibly is to imply that there are robust given mechanisms that allow for AI developments in ways that are ethical, can mitigate risk and promote good. It invokes governance and accountability and emerges from a complex ecology characterised by ongoing conceptual negotiation (ibid). Where responsibility lies, who has agency when it comes to responsible practices, who the actors are, and where dependencies lie are not necessarily linear, or mappable at the organisational level, resulting in descriptions of responsible AI as an ecosystem.

There are, however, factors that problematise the realisation of responsible AI. Social and cultural complexities mean that responsibility is likely to vary in accordance with local culture and politics (as in the case of Microsoft's Tay), the environmental cost of systems must be surfaced and understood so that users can make informed judgements about the true cost of use, and we must better understand the moral psychology of users and take steps to minimize the extent that companies can game and monetise our psychological responses. The distribution of power, often articulated in terms of available compute and access to data, needs to be more equitably distributed, and how organisations negotiate conflicting values and principles when adopting AI systems needs to be better understood. Finally, but perhaps most importantly, we must move beyond harm minimisation towards genuine opportunities for AI to enhance human flourishing and enable systemic diversity and sustainability.

3.6.1 Datasets, Bias and discrimination

AI innovation and progress is, in the main, driven by the quality and scale of training data (Longpre *et al.*, 2024). However, data collection practices are 'relatively immature', meaning that researchers tend to know little about the datasets used to train their models, including what they contain and

their provenance[48]. The efficacy of a model relies upon the scale and heterogeneity of data used to train it. Contemporary models are usually trained on data scraped indiscriminately from the Internet and this has become the most dominant source since 2019 (Longpre *et al.*, 2024; Birhane *et al.*, 2023). These data sources are controlled by large tech companies, most obviously Google, concentrating the power to train models within the hands of a small number of key players and, subsequently, reflecting their intentions and interests. A further byproduct of this, is that it has created a monoculture that does not accurately reflect the breadth of human nuance, behaviour and interest. Despite an increase in representation of non-western languages and geographies in the data, there is still an overwhelming 'western-centricity at an ecosystem level' (ibid). For example, western influences present even within predictive writing styles (Agarwal, 2025). This, coupled with the historical nature of the data used, results in an AI data ecosystem that is rich in bias and cultural and ideological influence, calling for greater inclusion of low resource languages[49].

The term bias refers to a statistical distortion within a dataset that can result in prejudicial outputs and outcomes. In the case of AI, the socially situated nature of that technology results in a greater likelihood that any undetected bias could result in discrimination and context-dependent harm. Sex, gender, disability and race-based bias (Criado-Perez, 2019; Hall & Ellis, 2023; Whittaker *et al.*, 2019), have been demonstrated within a wide range of AI applications, the most notable being facial recognition systems that fail to accurately identify race and gender (Buolamwini, 2024). Furthermore, it has been demonstrated that intersections of identities that face greater degrees of errors from facial recognition software create a cumulative effect that makes AI facial recognition even more inaccurate. This leads to, for example, older black women facing a significantly higher degree of bias from AI facial recognition (Sarridis *et al.*, 2023).

To take an example from a high-risk context, whilst AI has greatly enhanced digitized cancer imaging, image classification, detection and segmentation (detecting the tumour and nearby organs at risk), in the context of clinical oncology AI segmentation models have been shown to reflect sex-based bias in data generation, model building and clinical implementation due to the absence of sufficient female-specific data (Doo *et al.*, 2025), potentially negatively biasing health outcomes amongst women.

Bias can occur within many points in the AI workflow and, whilst we are most familiar with historical bias - where historical data does not reflect current conditions or norms such as the exclusion/omission of female-sex data within biomedical research - there are other forms which are equally important. For example, representation bias (where the sample doesn't reflect the target population) and measurement bias (where data is misclassified during curation). Whilst such issues remain a pressing concern, they are further compounded by the lack of representation within the AI workforce, which remains skewed towards white, male, north Americans due to a range of stable structural and cultural barriers (West *et al.*, 2019).

3.6.2 From bias to discrimination

Whilst bias simply describes a distortion in data, amplification of these distortions through AI models can result in discrimination, inequality and oppression, particularly when algorithms are used to categorise, predict and allocate resources (Eubanks, 2018; Noble, 2018), thus automating inequality. To give an economic example, algorithmic bias can result in some types of online content being suppressed by advertising-keyword blocklists, which are often maintained by third-party advertising agencies. Originally designed to suppress hate speech and objectionable content, in practice these systems have been shown to actively suppress advertising revenue

[48] https://thelivinglib.org/this-is-where-the-data-to-build-ai-comes-from/

[49] Low resource languages commonly refer to languages that are poorly represented in the linguistic datasets used for training in Natural Language Processing (NLP) tasks, e.g., a lack of speech data or annotated text. Other factors that determine whether a language is low resource are the local availability of human and digital resources, agency of the linguistic community members, availability of linguistic descriptions of a language and socio-political factors (Nigatu et al, 2024).

where articles reference queer or other content associated with marginalized groups. YouTube was called out specifically for their algorithm not featuring ads within videos that contain LGBTQ-related vocabulary, meaning that the creators see far less revenue than those developing more mainstream content (Romano, 2019). Whilst AI may eventually offer more sophisticated filtering, such issues remain. Content from marginalised groups is still, far too often, automatically associated with prohibited or unsafe content, resulting in already vulnerable groups losing out further (Kingsley *et al.*, 2022). With the recent dismantling of EDI programmes and initiatives in the United States, this problem is likely to drop in priority for platform providers.

The lack of transparency of more advanced AI models presents a yet higher risk, and the resulting concerns over accountability can leave those already socially and economically marginalized without clear pathways and instruments for recourse. Whilst regulations such as the EU AI Act offers some protection there remains a stark divide between the instruments designed for recourse and repair, and the capacity and capability of vulnerable individuals to make use of them. If we take policing as an example, historical bias inherent in police data means that any applications of predictive systems would almost certainly penalise those already vulnerable to institutional biases. For example, Richardson *et al.,* (2021) offer three cases of cities where dirty data (data resulting from 'flawed, racially biased, and sometimes unlawful practices and policies') has resulted in predictive policing systems that have been trained on inherently flawed data. In the cases of Chicago, New Orleans and Maricopa County 'dirty policing practiced', and a lack of public transparency in the latter location, enhanced the risk that those already racially marginalised within those areas would be at greater risk of flawed or unlawful predictions. In the cases where institutional change is most needed, biased predictive systems would be less likely to raise concern, and more likely to simply confirm and compound existing biases.

Case study: The Allegheny Family Screening Tool

One widely reported instance of algorithmic bias leading to discrimination comes from Allegheny County in the United States. The Allegheny Family Screening Tool was an algorithmic system designed to assess the risk level for children within a family, specifically in instances where child welfare concerns had been reported. The tool was conceived to support child-protective services in identifying whether a child might be at risk of abuse (anything from neglect to inadequate housing) and so predict whether a child was likely to be placed in foster care within two years of the initial investigation.

The tool was trained on a public dataset using "detailed personal data collected from birth, Medicaid, substance abuse, mental health, jail and probation records, among other government data sets, the algorithm calculates a risk score of 1 to 20: The higher the number, the greater the risk" (Ho & Burke, 2022). However, prior to the system being developed, analysis of the same data from that county was shown to include very specific human biases. For example, African-American children were referred three times more often than white children, despite "little evidence to suggest that their level of risk or need for services is substantially different than that of white children" (Rauktis & McCrae, 2010, p.4).

The family screening tool also reportedly tied negative outcomes to disability, compounding any existing biases in the system. This example represents an algorithmic system that, whilst clearly designed with good intentions, only served to reinforce years of systemic bias.

As this case illustrates, where AI systems are deployed in sensitive settings, the likelihood is that the complexity of social context will problematize their use and, where those systems are generative, mean that it will be impossible to identify, eradicate or control for those biases resulting in prejudicial outcomes. As described by Rauktis and McCrae in their research to document the service paths of African-American children in Allegheny:

> "Circumstances that are often experienced by African-American families, such as having a low income, living in an unsafe neighborhood, single parenting, lacking an education or using substances or having a serious mental illness were likely factors that make these families more vulnerable, increasing their visibility to systems such as child welfare. All of the interviewees felt that being poor and black were so intertwined that it was impossible to unravel them in order to determine which one caused African-American families to be disproportionately involved in the child welfare system". (Rauktis & McCrae, 2010, p. 5)

It is clear that the issue of bias within AI is inexplicable linked to fairness and justice and, as such, is of critical moral and social concern. Controlling for bias in datasets is a foundational step towards more ethical AI.

3.6.3 Bias Mitigation

Bias and, its mitigation, is one of the most researched areas of ethical concern in the context of AI. According to Sasseville *et al*., (2025), approaches to bias mitigation in health can be segmented into four categories: (1) the modification of AI datasets and models, (2) ensuring data come from robust data sources/electronic health records, (3) ensuring tools are developed with a human in the loop, and (4) through identification of a priori ethical principles to inform decision-making. Findings reflected in a prior review by Hall & Ellis (2023). Of these, the authors argue, it is 'algorithmic preprocessing methods' that show the greatest potential. In other words, ensuring that data is debiased is the most effective means by which bias might be managed, and involvement of multiple stakeholders at the pre-processing stage, as well as ensuring data are open-sourced, is critical to the minimisation and management of discrimination through AI (ibid). However, Bias mitigation efforts often create knock on effects that lead to the manifestation of new biases

elsewhere within machine learning systems; virtually all static bias mitigation efforts create knock on effects (Nizhnichenkov *et al*., 2023), which strengthens the case for algorithmic measures.

It must be noted that bias sensitivities are more likely to be experienced by those affected and so, without sufficient diversity and representation within datasets, the AI workforce and the wider AI governance infrastructure (Hall & Ellis, 2023), we are unlikely to see substantial progress on bias mitigation.

3.6.4 Synthetic data

The rapid scaling up of LLMs has meant that we are now questioning whether we have reached the practical limits of scalability. In the scientific journal Nature, Nicola Jones (2024) highlights research that flags the dwindling availability of conventional training data. This has resulted in a prediction that, by 2028, AI will have exhausted the estimated stock of public online text and so, essentially, run out of data. This comes at a time where content providers are also tightening restrictions and looking for enhanced copyright protections, forcing AI companies to seek partnerships with content providers (Jones, 2024).

In response to this and other access issues surrounding natural data (cost, availability, robustness), synthetic data is gaining traction as a means by which models might be trained. In mid 2024 NVIDIA announced[50] Nemotron-4 340B, a set of open models specifically designed to generate synthetic data for the training of LLMs. The outputs of these models mimic the characteristics of natural data and, whilst not yet in common use, have been experimentally shown to achieve what has been described as 'competitive performance' when used to improve LLMs, compared to models trained on human-annotated data (Sudalairaj *et al*., 2024), thereby potentially reducing reliance on human-annotations. The use of synthetic data has great potential in the context of health, where it could be used to supplement natural data by incorporating specific conditions and the kinds of edge cases not commonly found in such datasets, improving health outcomes for a wider range of people[51]. However, such methods must be used with care as the generation of synthetic data also holds the potential to further encode structural and historical biases. According to Hao *et al.,* (2024) such data can lack sufficient consideration of geographic and demographic diversity and draw biases from the models that generated them. The current lack of regulatory and ethical constraints on synthetic data further amplifies such concerns. Such datasets can also suffer from distribution bias, incompleteness, inaccuracy, insufficient noise levels, over-smoothing, the neglect of temporal and dynamic aspects found in natural data, and inconsistency (ibid) and so should be adopted with care. An alternative approach to the data scarcity issue, which is now gaining traction, is the development of smaller more efficient task-specific models, such as those developed by HuggingFace (Jones, 2024), though these are still at experimental stages. The unexpected efficacy, relative to cost, of DeepSeek has further galvanized this trend, as has the increasingly unpalatable energy costs of large models (ibid).

3.6.5 AI Literacy

AI literacy is core to contemporary democracy accountability. A citizen's ability to seek recourse if harm occurs, or to make reasoned and critical judgements about the information presented to them, is already a pressing need. Even before the mainstreaming of GenAI, research showed that individuals were mostly unable to distinguish human from AI-generated text (Kreps *et al.*, 2020; Clark *et.al*., 2021). However, the field of AI moves at pace but, despite there being a growing awareness of AI, public AI literacy is less advanced and does not extend to emerging technologies. For example, the Global Public Opinion on Artificial Intelligence survey (GPO-AI) explored opinion across 21 countries, finding attitudes to be varying and region-specific and whilst most respondents felt they knew what AI was (73%), this knowledge varied across AI applications. Only 30% or

[50] https://blogs.nvidia.com/blog/nemotron-4-synthetic-data-generation-llm-training/

[51] https://www.ibm.com/think/insights/ai-synthetic-data

respondents, for example, had heard of deepfakes (Loewen *et al.*, 2024) despite this form of AI being of core ethical concern.

AI literacy is more than general awareness of AI or simply a broad understanding of how AI works. As such systems permeate our lives, we will require skills that allow increasingly advanced critical evaluation of generated information and content including assessment of provenance, veracity, authenticity, and critical thinking skills more akin to media literacy than the Information and Communication Technology (ICT) skills of the past. Young people will need to develop the capability to evolve alongside the technology and in particular to "more explicitly include critical thinking so that they can exercise independent judgement over what is provided to them by GenAI/AI (or other sources)" (Meagher & Robertson, 2024). As with ICT, AI literacy will also require some form of citizen lifelong learning due the pace of innovation.

Long and Magerko (2020) argue that any AI literacy must be complemented and supported by AI interface design, supporting users to not only understanding how AI works but also reflect some aspects of interdisciplinarity, the strengths and weaknesses of AI, a focus on imagining future AI applications, understanding knowledge representation and decision-making, supporting explainability, understanding the human role in AI and data literacy, as well as AI ethics. By scaffolding more robust user mental models of AI, interface design can support informed user engagement with AI systems.

AI literacy will also be required at the organisational level if sectors are to be expected to engage in the AI ecosystem in meaningful ways. To this end Newman-Griffis (2025) suggests practice-based competencies that the author describes as 'AI thinking', shifting thinking from an innovation driven approach, where organisations seek to stay ahead of competitors, towards an approach that focuses instead on the specific goal or information-processing needs at hand. This approach, and others like it, call for a slowing of the charge towards innovation and instead focus on what the technology is for and whether it is in fact the best tool for the job at hand. AI literacy has a role to play in this more mindful approach to adoption. By focusing on understanding, and meaningful adoption as opposed to adoption from a place of fear of missing out, we might optimise for a broader range of human values than simply innovation. To date, AI innovation has focused on the optimisation of productivity and economic growth, but De Cremer and Kasparov (2022) argue that this perspective essentially narrows the potential of AI for good by overlooking the diversity of human interests and values. To this end AI designers and systems architects will also require skills development, in the form of moral awareness and training in responsibility (De Cremer & Kasparov, 2022). By equipping practitioners to consider responsibility at the earliest stages of the AI pipeline we move closer to mitigating some of the ethical problems further down the road.

3.7 Privacy, autonomy and consent

The datafication of human life has created a climate within which our understanding and expectation of privacy has been vastly altered. Unbroken surveillance (within both public and private spaces) has come to define most contemporary societies, and this has brought with it a global distribution of data-driven power that has placed unprecedented wealth and influence in the hands of a few organisations. Data has exploded in value at the same time as individual-level control over its use has eroded, partly through regulatory measures that allow its use, and partly through our growing dependency on the devices that generate and use that data. Once a by-product of human action data now not only drives how our past actions are revealed and future actions predicted, but it also better enables companies to capture our attention and influence our behavior.

Whereas once we spoke of privacy as boundary control, advances in AI have made it nearly impossible for an individual to exert any control over public/private boundaries as even acts such as turning off a computer or unplugging an internet connection are things of the past. Equally, the opacity of both the models and the wider data flows within the corporate AI ecosystem has rendered us functionally powerless through our dependence on the resulting products and services; "Today, it is basically impossible for people using online products or services to escape

systematic digital surveillance across most facets of life—and AI may make matters even worse" (King and Meinhardt, 2024). Therefore, how we protect our privacy remains an ongoing concern. To date, consent has been the primary mechanism by which we have protected our privacy and exercised autonomy online. However, as AI is built into the systems with which we interact, we are becoming increasingly decoupled from traditional computing devices and instead engaged in less explicitly visible computational interactions. In other words, as AI becomes more embedded into all areas of our lives, we are less aware or able to make privacy choices in real time, if at all. Researchers from the Stanford Centre for Human-Centred AI argue that three changes are required to redress the privacy imbalance:

1. A move away from 'data collection by default' by shifting from opt-out to opt-in data collection. This would involve a push towards data collectors meaningfully operationalising data minimisation through "privacy by default" strategies, coupled with adoption of technical standards/infrastructure to allow for meaningful user consent.
2. "Focus on the AI data supply chain to improve privacy and data protection. Ensuring dataset transparency and accountability across the entire life cycle".
3. Support the development of new governance mechanisms and technical infrastructure through policymaking so that the exercise of individual data rights and preferences is supported by design and by default.

(King and Meinhardt, 2024, p4)

Whilst instruments such as the EU ePrivacy directive and the EU GDPR – the global standards for data protection - have attempted this, such efforts have been functionally sabotaged by resulting design norms. For example, the a priori list of tick box consents that purposely build friction into the web-browsing experience in the hope that users will click away their rights as an irritation. This builds on practices known to result in a higher likelihood of user assent (Böhme & Köpsell, 2010).

It has been suggested that recasting consent as a social process, whilst reconnecting users with a sense of their data, could support stronger consent. Explicitly designing to support user comprehension, scaffolding their expectations through design-based affordances, offering regular feedback, and enhancing opportunities for interrogation and interaction (Luger & Rodden, 2013) would support both more meaningful consent and the growing AI transparency agenda. However, such practices are challenging of design norms as they both run counter to traditional usability design and are likely to disrupt corporate access to user data, thereby undermining monetisation.

3.7.1 Design, manipulation, and emotional AI

The potential of design in supporting both transparency and AI literacy has been largely overlooked within mainstream design practices and norms. Whilst there is considerable research in the area, this has yet to be translated into coherent design practice. Whilst AI may not have created online privacy issues, it has certainly compounded then, further enabling questionable practices such as targeted advertising, advert retargeting, and real-time bidding.

Capturing and maintaining a user's attention has become increasingly important in the highly competitive attention economy, keeping users engaged with your system whilst minimising distractions from other sources. However, attention is a finite resource and so several design strategies are employed.

Firstly, designers reduce cognitive load by ensuring tasks are simple and repetitive, allowing users to progress toward their goals with minimal effort, such as quickly clicking through terms and conditions without reading them. Secondly, they design clean, uncluttered interfaces to prevent irritation and distraction. Thirdly, they foster habituation by creating consistent interface patterns across systems, which essentially train users to navigate their platforms seamlessly, such as the use of ribbons across Windows applications. Finally, designers use techniques like hedonic adaptation to encourage users to return, offering frequent minor updates and small rewards that trigger a dopamine response. Such methods are widely accepted, legal, and form the backbone of

most online experiences without causing immediate harm (Luger, 2023). They do, however, make securing meaningful consent difficult and create experiences that are essentially addictive, resulting in users seeking engagement despite any potential costs to privacy. This addiction is a critical component of the data services model of monetisation upon which the Internet is built. The longer and more frequently a user engages with a product or service, the more data can be collected and the greater the profit.

Within this economy there is also a less neutral side to design. Dark design patterns deliberately manipulate users into making decisions including nudging users toward specific actions, exploiting cognitive biases through choice framing, employing clickbait and fake reviews, adding addictive features to keep users hooked, and using timers to create anxiety-driven decisions. Such practices, although often criticized, are prevalent in online spaces. This is the environment in which machine learning (ML) technologies are being integrated, underscoring the importance of context and application when assessing their impact (ibid), and whether the use of data is legitimate.

Data Brokers and Informed Consent

Data brokers share consumer and personal information that is purchased from third parties, or scraped from the internet, to companies for marketing and other purposes. Whilst the GDPR mandates that consent must be obtained before personal data is shared, data brokers often circumvent this requirement by collecting publicly available information or leveraging the consent users unknowingly give in license agreements. This creates a gap between the intended protection of consent requirements and actual practice, where individuals unknowingly accept broad data-sharing terms simply to access social media and other online services.

Essentially, data brokers fail to obtain informed consent and, in many cases, violate GDPR regulations (Ruschemeier, 2023). AI tools play a dual role in this privacy struggle—either aiding data brokers in scraping personal information for profit or assisting individuals in requesting data removal, as GDPR allows. However, as AI enhances the capabilities of data brokers, the need for stricter GDPR enforcement becomes more urgent. The intersec between such practices and vulnerable populations raise critical moral concerns.

3.8 Governance and accountability

The European Union AI Act is the first comprehensive AI regulation, foregrounding ethics and transparency and whilst the field of AI governance is new, it is quickly maturing with international principles such as the OECD AI principles and UNESCO's Recommendation on the Ethics of AI setting common ground (IAPP, 2024). In contrast, the UK recently (2025) framed AI as core to their national economic security and mentioned ethics only once in the context of unlocking national library data, instead opting to foreground assurance, regulation and security to mitigate risks[52], showing that national policy is not always in harmony with high-level global ideals. For AI to be developed and deployed responsibly, it must be robustly governed, ideally with some form of global consensus as the platforms making use of AI operate across borders. Governance is a broadly used but variously defined term, most to describe international cooperation outside the state system (e.g. organisations such as the EU), proper implementation of state policy through public administration, or the regulation of human behaviour through non-hierarchical systems such as civil society bodies and actors (Fukuyama, 2016), incorporating non-state actors across the global political ecosystem.

Since 2016, the drive towards effective governance of AI has seen a raft of emerging principles, laws and regulations, AI frameworks, declarations, voluntary commitments and standards (IAPP, 2024). Initially driven by the public sector (Schiff *et al.*, 2020), efforts are now widespread, though they have also seen contest, being framed as either insufficiently robust or, at the opposing end of

[52] https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

the spectrum, limiting innovation, with corporate voices developing their own frameworks and calling for a shift towards self-regulation. The subsequent rise of AI industry self-governance has been variously described as ethics washing or ethics theatre, implying that such initiatives could serve to sanitise ongoing poor practice and allow corporations to cherry pick the measures they put in place (Wagner, 2018).

There have been numerous efforts to develop technical solutions in support of AI governance, such as the development of tools and techniques for fairness/bias mitigation (Johnson 2024). Whilst such tools can mitigate some issues, there is an ongoing trade-off between the desired outcome (e.g., fairness) and the subsequent quality of the output (ibid). Whilst we have seen some progress in the translation of ethical principles into AI practice, there remains a considerable gap (Sanderson, 2024) with the required trade-offs seemingly unpalatable in many cases. In other words, aligning AI innovation to human values necessarily requires some curb on unfettered speed, scale, cost or deployment of AI. However, how organisations manage such trade-offs remains under-researched. To this end Johnson (2024) suggest a three-stage process that involves proactively identifying trade-offs, prioritising and weighting them, and finally justifying and documenting the decisions made to support transparency and accountability.

However, fully understanding the sociotechnical to the extent that one might operationalise human values within the AI lifecycle (planning, design, development and deployment) requires a broadening of the disciplines involved in that process, and the development of tools and methodologies to enable effective cross-disciplinary working. Equally, articulating human values requires proactive commitment from all key players and a common understanding of their articulation, something currently absent at the global level. For example, whilst we might all agree to free speech as a value, its current articulation through the policies of Twitter/X and Meta are unlikely to reflect a global understanding of how that should be operationalised. This worrying trend, of shifting responsibility for both content and self-protection to the user, opens the floodgates to further mis/disinformation and compounds the need for global AI literacy.

3.8.1 Embedding the social sciences and humanities

Ensuring AI governance is effective within a fast-moving landscape is an ongoing challenge, but some factors enhance effectiveness of efforts such as strong links to regulation, specificity, reach, enforceability and monitoring, and iteration and follow-up (Schiff *et al.*, 2020, p. 157-158). Equally, Oduro and Kneese (2024) argue that the development of effective governance requires looking beyond technical solutions and towards social sciences and humanities knowledge.

> "Early research has shown that when sociotechnical approaches are integrated into AI development and testing and in use-feedback, positive outcomes significantly increase for impacted communities, users, and AI developers. Sociotechnical research and approaches have proven crucial to AI development and accountability — the key will be implementing AI governance practices that employ the expertise required to reap these benefits." (Oduro and Kneese, 2024, p.1)

The authors highlight a range of methods that demonstrate the importance of social sciences and humanities knowledge in AI governance such as auditing and impact assessments, GenAI assessments, and methods of public participation. They also recommend (a) investment in social sciences and humanities experts, (b) mandated incorporation of sociotechnical research and evaluation methods to inform procurement processes, and (c) inclusion of social sciences and humanities expertise in the development of all standards and guidelines for AI assessment, research and development and policy (ibid, p.5) to ensure such knowledge is embedded.

Despite this, it is clear that the global AI ecosystem has been driven by a technocentric and deterministic view that prioritises technical over social and humane progress. This orientation has become further entrenched through the second Trump presidency in the United States, where AI innovation has become an arms race, as illustrated by the $500 billion Stargate project and the abandonment of commitments to safe and responsible AI (Wilkinson, 2025) in a drive to maintain global leadership. More broadly, global political uncertainty and polarisation has reinforced the

state system and its focus on national security, including AI. For example, in the UK, the national AI Safety Institute was recently rebranded as the AI Security Institute, removing any focus societal harms and safety, to work more squarely on security (Meyer, 2025). At the same time, a disproportionality between Science, Technology, Engineering and Mathematics (STEM) funding has far outstripped that of the Arts and Humanities with this divide growing ever larger (Newfield, 2025) prioritising specific forms of knowledge. Despite this trend, there remains a real need for the arts and humanities to underpin responsible AI innovation so that AI works for, rather than against, social innovation. There are, however, some new programmes that are beginning to seek to address this imbalance both at the ecosystemic level and AI workflow levels. Schmidt Sciences, the philanthropic initiative of Eric and Wendy Schmidt, recently launched the first call of their newly launched Humanities and AI Virtual Institute, seeking to fund humanities research focused on specific technical issues[53]. In the UK, the Arts and Humanities Research Council (AHRC) launched BRAID in 2022 with a view to embed Arts and Humanities research into the wider responsible AI ecosystem.

[53] https://www.schmidtsciences.org/humanities-and-ai-virtual-institute/

Case Study: Embedding Arts and Humanities knowledge into a national AI Ecosystem (BRAID UK)

Bridging Responsible AI Divides (BRAID) is a £15.9 million UK-wide programme that integrates Arts and Humanities knowledge and research more fully into the AI ecosystem, as well as bridging the divides between academic, industry, policy and regulatory work on responsible AI. The programme is funded by the Arts and Humanities Research Council within UK Research and Innovation (UKRI), the national funding agency investing in science and research in the UK.

BRAID is a six-year programme (2022-28) designed to promote and enable responsible AI in the UK. The programme works in close partnership with the Ada Lovelace Institute[54], an 'independent research institute with a mission to ensure that data and AI work for people and society', and British Broadcasting Corporation (BBC) Research and Development[55]. The programme seeks to build a more disciplinarily integrated community working towards responsible AI innovation. It does this by supporting a network of interdisciplinary researchers and partnering organisations through the delivery of funding calls, community building events, and wider programmed research and activities[56]. Each of the projects funded through BRAID support scholars from the Arts and Humanities and embeds them with non-academic stakeholders to a clear pathway to knowledge exchange, integration and impact. Each project focuses on a specific area of concern and either seeks to understand the scope of the issue, demonstrate how that issue might be solved, or both. Scoping projects[57] have involved topics such as understanding how to assure and make trustworthy digital twins, embedding ethical review to support responsible AI in policing, respectful management of indigenous knowledge in the context of AI, creating responsive assessment of AI, understanding public engagement with AI in everyday life, embedding responsible AI in the school system, use of creative AI, support for human-AI collaboration, and use of AI to communicate colonial museum collections. One of three new, larger scale, demonstrator projects will explore the sustainability and environmental resilience of using AI, with two further projects looking at AI within the creative industries[58]; one of the most disrupted sectors. The programme is also supporting 17 fellows directly[59] and a further 7 fellows in partnership with the British Academy, the latter of which will find each scholar embedded within a different government department in the UK. The final three years of BRAID (2025-28) will extend this work to focus on critical AI literacy, public media and democracy, and societal resilience and repair[60].

3.8.3 Accountability and AI

Accountability is the cornerstone of effective governance but its context-dependence, the sociotechnical nature of AI systems, and the ambiguity of political processes (Novelli, 2024, p.1871) mean that ensuring accountability of AI is currently just out of reach. Accountability can be defined as "an obligation to inform about and justify one's conduct to an authority" (Novelli *et al.*, 2024, p.1872) and, in the context of AI, can best be understood as answerability which requires a recognised authority, the ability to interrogate a system, and a limitation of power. The requirement that AI systems be answerable is further problematized by the rise of GenAI as it involves an artificial agent. AI accountability requires that those responsible for AI systems will comply with

[54] https://www.adalovelaceinstitute.org/news/ai-opportunities-action-plan/
[55] https://www.bbc.co.uk/rd
[56] https://braiduk.org/scoping-call-projects
[57] https://braiduk.org/scoping-call-projects
[58] https://braiduk.org/demonstrator-projects

[59] https://braiduk.org/projects-fellowships
[60] https://braiduk.org

legislation and standards to ensure their products function properly during their lifecycle and in accordance with those constraints (ibid). Novelli *et al.,* (2024) argue that this view is overly limited in that it does not consider the broader sociotechnical context such as the history of training data or those affected by the system when deployed. In this way AI accountability is framed as a complex and multifaceted concept requiring broad ecosystemic and sociotechnical understanding and the views of multiple stakeholders.

3.8.4 Multistakeholder engagement and coproduction

The global AI ecosystem is cross-sectoral, complex and diffuse often fracturing along sectoral and disciplinary lines. Effective AI governance requires both multistakeholder involvement and robust mechanisms for collaboration and coproduction to ensure a balanced regulatory environment (Kaveh & Eisenberg, 2023). Multistakeholder collaboration is also more likely to result in trust, workforce transformation, mechanisms for accountability, and stimulation of organisational self-governance (Vincenzi *et al.*, 2024). Such approaches can also support public awareness and education (Kaveh & Eisenberg, 2023) and internal workforce transformation (World Economic Forum, 2024). The value of openness must also be a core characteristic of such initiatives as evidenced by examples of emergent practice.

WeBuildAI is a participatory framework that seeks to involve lay users directly in the design of algorithmic governance policy through a guided approach that allows non-expert communities to be directly involved in the design of the AI model itself (Lee *et al.*, 2019). Such newly emerging AI Collaboratives focus on enhancing diversity in the communities that build, regulate and deploy AI (Ding *et al.*, 2023). Whether approaches designed for lay user coproduction scale is yet to be tested, though there are scaled examples of expert communities outside of mainstream big tech. Three notable nongovernmental examples are: (1) BigScience workshop[61], an open science project supported by Huggingface[62] comprising hundreds of researchers globally who have collaboratively developed a large Open Multilingual Language Model (BLOOM); (2) The Turing Way[63], an open science community, supported by the Alan Turing Institute, guided by a 'reproducible, ethical and collaborative data science' handbook where resources and materials are shared, and (3) Mozilla's Trustworthy AI working groups[64] where AI builders and civil society actors can propose working groups on a topic of interest, the rules being the work should be experimental, deliverable and open licence (Ding *et.al.,* 2023). A further United States governmental example is the Artificial Intelligence Safety Institute Consortium (AISIC)[65], a consortium of 280 organisations that collaborate on measures, methods, criteria and models to support safe and trustworthy AI innovation. An example closer to home is offered by Transkribus, a cooperative seeded by EU funding and further developed by the community it was designed for.

[61] https://bigscience.huggingface.co

[62] https://bigscience.huggingface.co

[63] https://huggingface.co

[64] https://book.the-turing-way.org

[65] https://www.nist.gov/aisi/artificial-intelligence-safety-institute-consortium-aisic

Case Study: Collaborative open-source AI

READ-COOP is a non-profit cooperative that aims to make historical documents accessible by democratising text-recognition technology through their AI platform Transkribus. Emerging from projects funded by the European Union ('Transcriptorium') and READ (Recognition and Enrichment of Archival Documents) it responded to a genuine need from the academic and heritage communities and has successfully automated recognition of hand-written historical documents. In 2019 "more than 20,000 people from around the world, from countless fields and with innumerable backgrounds, had signed up, all with the same goal: to make the contents of historical documents accessible".[66] In 2020 it won the European Union's Horizon award for Impact. It now has 300,000 users worldwide and has processed over 100m images of historical text (Terras *et al*, 2025). The development of Transkribus took a 'bottom-up' approach, driven by motivated researchers, from multiple disciplines, who were willing to share and exchange user transcriptions and models to build and improve the corpus (Nockels, 2022). The tool is now reportedly the "most popular user-facing platform for producing transcripts of historical texts across the cultural and heritage industries" (Nockels et al, 2022), with evidence of it being applied in areas as distant as architecture and botany (ibid). Initially free at the point of use, Transkribus has since moved to become the cooperative-based paid-for model (2020) as their grant funding ended in 2019, to ensure sustainability of the service. The platform is now co-owned by over 250 individuals, institutions, and companies, and all revenue is directed back into the cooperative to support the platform's continuation and further development. The ongoing use of, and community investment in, this system illustrates the utility of AI when it is designed from the bottom-up, by and for communities of practice.

A further example from Europe is AI4Europe, a European Union initiative aimed at advancing AI across Europe. It provides a collaborative platform offering AI tools, resources, and models to researchers, businesses, and public institutions. The project focuses on ethical AI development, ensuring compliance with European standards, and supports small and medium-sized enterprises (SMEs) in integrating AI. AI4Europe also offers educational resources to build AI skills across the continent[67]. These examples illustrate the efficacy of cooperative action, but also demonstrate the critical need for seed funding to establish the tools and conditions for such communities to take advantage of AI.

[66] https://readcoop.org/why-coop
[67] https://www.ai4europe.eu/

Case Study: Involving the Voice of Children in AI Policymaking (The Children and AI Project)

Children and young people represent some of the key voices missing within current discourse around AI innovation, even though this class of technologies will disproportionately influence their lives when compared to prior generations. The Scottish AI Alliance[30] have tackled this challenge head-on through a partnership, between the Children's Parliament and the Alan Turing Institute in the UK, through the Children and AI project.[31] The project progresses the Scottish AI Strategy's commitment to adopt UNICEF's policy guidance on children and AI, and Scotland's broader commitment to the United Nations Convention on Rights of a Child (UNCRC).[32]

By foregrounding children's rights the project asks, 'what needs to happen for AI to play a role in keeping all children happy, healthy, and safe?' To answer this question, they began by exploring children's views on AI through a survey and a series of workshops in schools, introducing children (7-11 years) to both AI and their wider rights, exploring how these might intersect. This direct deliberation resulted in the development of four categories of focus (fairness & bias, learning about AI, AI in education and the Future of AI). Throughout the process, children were involved as experts, uniquely able to represent their views. The final report from the first stage can be found here.[33]

Within the second phase, again working directly with children but this time partnering with organisations working on specific AI policies or projects that the children can influence and then working with creative practitioners to express their learnings, the children developed twelve 'calls to action'[34] on the basis of their findings. These twelve calls included ensuring inclusion, taking steps to avoid any negative impact of AI, the need to involve children in the development of AI, the requirement that AI does not undermine children's rights (such as the right to appropriate and accurate information), opt-in consent to data use, AI as supporting rather than undermining human teachers, and embedding AI literacy in the curriculum[35]. These ideas, emerging directly from children, not only reflect existing thinking around responsible AI, but go further in pushing for a more normative climate for AI development. The third and closing phase is currently underway (March, 2025) with the aim to develop a series of resources to operationalise the learnings from the project. Outputs will include resources for AI professionals around children's rights and how AI impacts them, as well as resources about AI for children by children.

The resulting children's calls to action have been presented to decision makers across the policy landscape including local and national government, and educational bodies. The resulting resources will be made widely available to the public and private sector to ensure that children's voices are considered in decisions around AI in Scotland. The project hopes to raise awareness of the importance of children's perspectives in the development and deployment of AI and has been recognised as an exemplar in children's engagement with AI.

3.9 AI, Wealth Distribution and Inequality

3.9.1 Applications and impacts of AI across the sectors

The integration of artificial intelligence (AI) into the global workforce can be largely attributed to its adaptability and wide-ranging applicability across various industries. AI technologies are increasingly being used to automate tasks, improve efficiency, and enhance decision-making processes. Whilst the degree of AI adoption differs across sectors, its impact is becoming evident even in industries within the primary sector, including agriculture, mining, and logging. These

industries, which traditionally rely heavily on manual labour and natural resource management, are gradually incorporating AI to improve operations and increase productivity.

AI affects work everywhere

Artificial Intelligence (AI) is transforming workplaces globally, promising unprecedented efficiency, innovation, and opportunities for growth. However, the integration of AI also comes with challenges. The expected scale of AI integration into workplaces is enormous—about 40% of workers worldwide are in occupations with high exposure to AI, and this figure rises to 60% for workers in advanced economies (Cazzaniga *et.al*., 2024). AI is already widespread in economies with advanced digital infrastructures, and as AI technologies mature and integration continues, it will by no means remain a niche phenomenon. In the *IMF Staff Discussion Note Gen-AI: Artificial Intelligence and the Future of Work*, a framework for understanding the interaction between AI integration and work is laid out through the lenses of exposure—the degree to which AI can perform the tasks within a profession—and complementarity—the degree to which AI assists human productivity instead of replacing it. Occupations with high exposure to AI exhibit both high and low levels of complementarity. Those with high complementarity are expected to witness significant productivity gains without necessarily facing job insecurity, whilst those with low complementarity are at the greatest risk of AI-driven automation. The degree of complementarity and exposure varies by country, occupation, and sector of the economy. It is generally evident that advanced economies and the global tertiary sector will experience the most disruption, but they will also harness the greatest benefits from AI integration (Cazzaniga *et.al*., 2024). The anticipated widespread diffusion of AI across sectors and economies worldwide may have significant implications not only for the labour market but also for the broader economy.

3.9.1.1 The Primary Sector

In agriculture, AI-driven precision farming techniques are being employed to optimize crop production and reduce resource usage. For example, AI has been shown to reduce water consumption by as much as 25%, a significant step towards more sustainable agricultural practices (Rashid & Kausik, 2024). Farmers are using AI systems to monitor crop health, manage irrigation, and analyse soil conditions, enabling more informed decisions. Additionally, AI-augmented drones are being used for tasks such as detecting diseases and weeds, whilst autonomous tractors are automating essential farming activities like ploughing and planting. In the mining industry, AI is being used for tasks like data analysis during prospecting, helping identify mineral deposits more accurately. Autonomous vehicles are also beginning to play a role in streamlining mining operations. Whilst AI applications in mining are still at an early stage, they are steadily gaining traction and contributing to improved efficiency. Logging has seen less direct application of AI technologies; however, advancements in AI for environmental conservation, such as monitoring forest health and ecosystem management, suggest potential areas of crossover that could benefit the logging industry. Employment projections for software developers indicate that there will be increases in employment across the primary sector, showing that AI will continue to be integrated in these industries (Chan, 2024; Rashid & Kausik, 2024).

Whilst AI applications in the primary sector are widespread and harbour potential for economic disruption, they remain primarily focused on complementing the existing workforce and increasing productivity rather than replacing human labour. This is due to the nature of the tasks in these industries, which often require a combination of manual and technical expertise. Decline in employment in professions such as agriculture will most likely continue to be driven by automation in general (UNFAO, 2022), rather specifically AI driven automation. In contrast, the secondary and especially the tertiary sector are expected to experience a more significant impact from AI technologies. These sectors, which involve manufacturing, logistics, and services, provide a broader scope for AI to transform processes, optimize workflows, and enhance customer

interactions. As a result, the influence of AI on the global workforce is likely to expand further, with its greatest impact observed in industries with more diverse and complex operational needs.

3.9.1.2 The Secondary Sector

AI plays an increasingly vital role in the secondary sector of the economy, where it is transforming traditional processes and enabling new efficiencies. In planning and logistics, AI tools are employed to optimize workflows, manage supply chains, and enhance resource allocation. The design and creation of goods also benefit from AI technologies, which can produce multiple designs based on prompts and optimise the selection of materials. In the construction industry, AI-driven tools help reduce waste and optimize resource use (Chan, 2024). AI processes sensor data and provides predictive analytics for maintenance, supply chain management and other applications. AI also enhances quality assurance through advanced detection systems that identify defective products in assembly lines, ensuring that only high-quality goods reach the market (Chan, 2024). These applications not only improve productivity but also contribute to reduced environmental impact and operational costs.

In manufacturing, AI techniques such as Artificial Neural Networks (ANNs) are used to model complex industrial processes, enabling precise control and prediction (Rashid & Kausik, 2024). These networks are particularly valuable in applications where systems must operate under specific constraints, such as energy consumption or structural durability. Additionally, AI-based global optimisation methods, including Genetic Algorithms, Evolutionary Algorithms, Particle Swarm Optimisation, and Ant Colony Optimisation, are applied to solve multifaceted challenges like minimising production costs, maximising structural stiffness, and optimising the efficiency of manufacturing pathways (Rashid & Kausik, 2024). These techniques allow manufacturers to address competing priorities whilst maintaining high performance.

According to the European Parliamentary Research Service (EPRS), AI is at the heart of Industry 4.0, the digital transformation reshaping industrial production. Future factories are envisioned as spaces where physical and digital systems merge seamlessly, with AI driving data analysis, process optimisation, and maintenance automation. By leveraging AI, manufacturers can boost quality control, minimize downtime, and achieve continuous improvements, ultimately being the catalyst for the creation of industries which do not yet exist (Marcin, 2019).

AI driven robotics are expected to automate repetitive, high volume and labour-intensive processes in industrial work, potentially freeing up time for operators to do more sophisticated work (Rashid & Kausik, 2024), however, labour optimisations could eventually also lead to less personnel being needed in certain tasks. The advancement of AI robotics, optimised resource allocation in construction and the automation of many time-consuming clerical, logistical and administrative duties could lead to job loss. For example, It is estimated that 38-45% of current jobs in construction will be handled by AI analytics and robotics by 2030 (Chan, 2024; Regona 2022). Employees will likely have to work alongside robots that handle specific aspects of their work, skilled workers are unlikely to be replaced to avoid the loss of expertise, and new jobs will almost certainly be created, such as AI engineers, Technicians and Trainers (Chan, 2024). The secondary sector will still experience employment growth worldwide, but the skills required as a barrier of entry are likely to increase, as simpler, repetitive or low skill work may face automation. Rising skill requirements will likely cause greater disruptions for employment in developed economies with more substantial digital infrastructure (Chan, 2024). The European Centre for the Development of Vocational Training (CEDEFOP) employment growth forecasts in the EU27 for 2022-2035 predict negative employment growth for clerical workers, trades and manufacturing workers, despite sectoral growth and rising worker productivity (CEDEFOP, 2020; Eurostat, 2024). These trends are indicative of automation driving low skill workers out of the secondary sector, whilst boosting the productivity of the high skill workers that remain. AI is a major driver of further automation and rising skill requirements in already mature industries such as manufacturing and construction, so further disruption to labour in the sector is to be expected (Regona 2022; Tyson & Zysman, 2022).

3.9.1.3 The Tertiary Sector

The tertiary sector is a vast and diverse area of the economy encompassing thousands of occupations. These jobs require a wide array of skills and involve varying levels of human interaction. Activities in this sector range from research, financial services, and coding to emergency services, education, retail, customer support and many more varied occupations. As the sector where most people are employed and the foundation of much of the global economy, it is particularly important to understand its vulnerability to the transformative potential and wide applicability of AI in occupations therein. AI has the ability not only to enhance productivity and streamline processes but also to fundamentally alter or eliminate certain roles and occupations within the sector (Cazzaniga et al., 2024).

Historically, advancements in automation have primarily impacted the primary (agriculture and resource extraction) and secondary (manufacturing and production) sectors. This has often led to significant portions of the workforce transitioning to the tertiary sector. The service sector has traditionally offered resilience to automation because of the unique human qualities required in many of its roles, including insight, creativity, and interpersonal interaction. However, with the rapid advancement of AI, even this historically resilient sector is facing unprecedented levels of disruption (Tyson & Zysman, 2022).

AI has already been implemented across various occupations in the tertiary sector, automating tasks that were once considered safe from technological disruption. Examples of these changes are abundant and indicative of broader trends within the sector. For instance, stores like Amazon Go have replaced cashiers with AI-driven systems that use cameras and sensors to monitor purchases and automate checkout processes. Similarly, algorithms have played a critical role in content recommendations on large social media platforms for years, personalising user experiences whilst reducing the need for human curation (Chan, 2024).

In education, AI has been utilized to assist educators in numerous ways. It is used to score tests, predict student outcomes, and even tutor students in specific subjects. AI systems can help students manage applications, meet deadlines, and learn programming, effectively taking on roles traditionally performed by academic advisors (Rashid & Kausik, 2024).

Marketing is another domain where AI has made significant inroads. Through targeted advertising, AI algorithms have revolutionized how businesses reach their audiences. AI-generated language and art now support marketing campaigns, potentially reducing the need for highly skilled creative professionals (Rashid & Kausik, 2024). Creative professions face multiple risks from AI art; beyond advertising, AI art threatens the livelihood of creatives and creativity in general. AI-driven marketing strategies allow companies to rely on AI tools rather than human expertise, this was demonstrated by the example of the fintech firm Klarna, which used AI to generate thousands of images corresponding to major marketing events, and drive larger campaigns whilst reducing its costs by $10 million (Mukherjee, 2024; Caporusso, 2023).

This cost-cutting measure allows companies to rely on AI-driven marketing strategies rather than human expertise, as demonstrated by the example of Klarna, which used AI to generate thousands of images corresponding to major marketing events, and drive larger campaigns whilst reducing its costs by $10 million (Mukherjee, 2024).

Retail giant Amazon showcases AI's versatility in services such as home price estimation, facial recognition, and autonomous driving. In the banking sector, AI is widely employed to detect fraudulent activities and interact with customers. (Rashid & Kausik, 2024). Additionally, large financial firms use AI for investment advice and decision-making processes (Miziołek, 2021). Many customer service roles have been replaced or supplemented by AI chatbots that handle queries instead of human workers in call centres. Heathrow airport for example uses AI driven holograms to answer customer queries, these robots and others like them can work 24/7 with little cost, and since they are not human, they do not experience fatigue from emotional labour or mundane and repetitive tasks (Wirtz & Pitardi, 2023). Whilst this frees up human labour for more creative tasks, there is no guarantee that all workers will be able to transition to such work or be retained to do so (Cazzaniga *et.al*., 2024).

In technical fields like software development, AI is automating substantial portions of coding work, thereby streamlining processes, increasing efficiency and reducing the work needed to be done by human coders, driving fears of skill loss and automation (Kakhiani, 2024). Clerical and administrative tasks, once the backbone of many office jobs, are also being automated or are at high risk of being automated by AI in the future, as the need for human intervention in repetitive administrative activities is reduced. (Gmyrek, 2023)

The widespread adoption of AI in the tertiary sector is expected to result in significant disruptions. Workers in roles with high exposure to automation but low complementarity to AI are likely to transition to other jobs. This shift is expected to polarize the workforce. On one end, highly skilled and managerial professionals will adapt to and leverage AI to their advantage, enhancing their productivity and value. On the other end, lower-skilled, lower-paid workers will be retained for roles where human-to-human interaction and trust are indispensable. (Chan, 2024; Cazzaniga *et.al*., 2024)

Middle-skilled workers, however, are at considerable risk. Those engaged in repetitive or intensive tasks that can be automated, yet lack the ability to transition to more cognitively demanding roles, may face significant challenges in maintaining employment. Older workers, who might find it difficult to adapt to new technologies or transition to different occupations, are similarly vulnerable to job displacement (Cazzaniga *et.al*., 2024).

The impact of AI on the tertiary sector is profound and multifaceted. Whilst it offers opportunities to enhance efficiency and innovation, it also presents challenges in the form of workforce disruptions and job losses. The future of the sector will depend on how workers, businesses, and policymakers respond to these changes, ensuring that the benefits of AI are balanced with strategies to mitigate its adverse effects on employment and society.

Case Study: Digital Switzerland Strategy 2023: Collaboration and ethics in an anthropocentric approach

Switzerland's Digital Strategy 2023 serves as a model for reducing resource disparities in AI adoption by emphasising employee education, enacting digital-friendly legislation, and fostering cross-sector collaboration. This holistic approach ensures that businesses of all sizes can leverage AI-driven growth whilst maintaining ethical and regulatory standards.

To enhance digital literacy, Switzerland invests in several educational initiatives. The Federal Data Science Strategy (2022) promotes AI awareness and competence among professionals, ensuring that employees across industries develop the necessary skills for digital transformation. Platforms like Renku provide open-access AI research tools, fostering knowledge-sharing and innovation. Additionally, Switzerland's world-renowned academic institutions, such as ETH Zurich and EPFL Lausanne, play a crucial role in AI research and talent development. The ETH AI Center, for instance, encourages interdisciplinary collaboration between business, politics, and society, reinforcing Switzerland's commitment to widespread digital education. Switzerland also maintains a flexible legal framework to encourage AI adoption whilst safeguarding public interest. The revised Federal Act on Data Protection (FADP) 2023 ensures transparency in AI-driven decision-making, particularly in cases involving personal data. Instead of implementing a single rigid AI law, Switzerland tailors its regulatory approach to specific industries, allowing businesses to innovate whilst maintaining consumer protection. This sector-specific approach prevents excessive legal burdens on smaller enterprises and startups, creating a level playing field in AI-driven markets. Public-private partnerships further drive AI development by encouraging knowledge-sharing and investment. Innovation Sandboxes, such as Zurich's AI pilot projects, allow businesses, academia, and policymakers to experiment with AI applications in a controlled environment, facilitating research commercialisation. AI policy networks also connect businesses, research institutions, and government agencies to establish best practices and enhance sectoral collaboration.

Switzerland's inclusive digital strategy levels the playing field for AI adoption, ensuring that small and large enterprises alike can benefit from digital transformation. By prioritising education, supportive legislation, and collaboration, Switzerland provides a sustainable and ethical framework for AI integration, making it a global benchmark for responsible digital innovation. (Digital Switzerland Strategy 2023; AI Watch; Fedlex, 2020; Deloit, 2023)[68].

3.9.2 The Economic Effects of AI: Productivity, Growth, and Inequality

Artificial Intelligence (AI) has far-reaching implications for the economy, touching on productivity, income distribution, and the structural organisation of labour. As the applications of AI continue to expand, its impact becomes increasingly multifaceted, presenting both significant opportunities and profound challenges (Eloundou, 2023). Whilst the positive effects of AI can inform us as to

[68] https://ai-watch.ec.europa.eu/countries/switzerland/switzerland-ai-strategy-report_en#infrastructure;https://digital.swiss/en/strategy/strategie.html; https://www.fedlex.admin.ch/eli/cc/2020/988/de;https://www.deloitte.com/ch/en/Industries/government-public/perspectives/an-overview-of-ai-in-the-swiss-public-sector.html

why it has the potential to drive economic inequality, it is essential to contextualize these dynamics to understand their broader implications.

3.9.2.1 Productivity Gains and Economic Growth

AI's ability to automate tasks and enhance human productivity has become a cornerstone of its economic impact. By handling routine, repetitive, or highly complex workloads, AI enables businesses to optimize operations and improve outcomes. For workers, AI offers tools to tackle larger and more intricate tasks, amplifying their efficiency and effectiveness.

Given its cross-industry applicability, AI is poised to drive productivity gains across a majority of sectors. Larger firms, which often have the resources and infrastructure to adopt AI early, are likely to see these gains reflected in increased national incomes and wages for employees who successfully leverage AI tools. The integration of AI into routine business practices can streamline operations, reduce costs, and create efficiencies that contribute to broader economic growth.

3.9.2.2 New Sectors and Opportunities

The transformative potential of AI extends to the creation of entirely new sectors dedicated to its development, maintenance, and application. From data science and machine learning engineering to AI ethics consulting and cybersecurity, the industries surrounding AI are likely to employ a significant number of highly skilled professionals. These sectors not only generate economic activity but also create opportunities for innovation and specialisation.

However, the benefits of these developments are not universally accessible. The demand for expertise and the high barriers to entry in AI-related fields can limit participation to a relatively small, highly skilled workforce. This creates a dichotomy in which certain groups thrive, whilst others face stagnation or decline (Karippacheril, 2024). Further, sub-economies based around the outsourcing of AI related tasks may emerge, such as the creation of data annotation work centres in places like India, on the one hand contributing some economic growth in poorer countries, whilst on the other, working to benefit the outsourcers much more.

3.9.3 Artificial Intelligence and Inequality across the board

Whilst the benefits of artificial intelligence (AI) are significant, its integration into the economy raises critical questions about inequality and its disruptive effects on social structures.

Historically, technological revolutions—from industrialisation to the digital age—have transformed economies and societies whilst also deepening inequalities. The First Industrial Revolution transformed an agrarian society into a factory-based one, creating a two-tier economy and a deeply unequal society. As Karl Marx observes: "....accumulation of wealth at one pole is, therefore, at the same time accumulation of misery, agony of toil slavery, ignorance, brutality, mental degradation, at the opposite pole..." (Karl Marx, Capital. Volume 1, Chapter 25) [69]. The Second Industrial Revolution concentrated wealth among industrialists whilst exploiting workers and creating wage gaps based on gender and race. Men were considered breadwinners, women were supplementary. The Third Industrial Revolution, despite progress and advancements, with automation and digitalisation, widened economic disparities, weakened job security, and fuelled corporate dominance.

[69] https://www.marxists.org/archive/marx/works/1867-c1/ch25.htm#S4

Case study: *Thank You, Mrs. Mary Tsingou*: Human Computers and systemic inequalities

In 1955, the Los Alamos Scientific Laboratory published the paper *Studies of Nonlinear Problems*, detailing the methods and results of a mathematical physics simulation run on the MANIAC, the laboratory's first electronic computer. The paper, authored by Enrico Fermi, John Pasta, and Stanislaw Ulam, was immediately recognized for its groundbreaking simulation. Today, this simulation is known as the Fermi-Pasta-Ulam (FPU) problem. In a footnote, the authors acknowledged, "W*e thank Miss Mary Tsingou for efficient coding of the problems and for running the computations on the Los Alamos MANIAC machine*."

Mary Tsingou was born in 1928 in Milwaukee to Greek immigrant parents. She studied mathematics but faced a difficult job market for female math teachers. In 1952, she applied for a position as a mathematician at the Los Alamos Laboratory. Two years later, in 1954, she returned to university to earn a master's degree. During the Korean War, the laboratory sought female mathematicians, as men could be drafted at any time. Upon being hired, Tsingou and other young women were informed they would be paid less than their male counterparts, despite having the same skills and qualifications, because *"men were breadwinners and women were just supplementary.*" Tsingou soon became one of the first programmers for the MANIAC (Mathematical Analyzer, Numerical Integrator, and Computer). "*I was very interested in learning programming,*" she recalled in a recent interview, "*because it was pretty boring sitting there doing addition and subtraction.*" However, she noted, "t*he men always got the more interesting problems, and the women were always relegated to the mundane—keeping the machine going and stuff like that.*" [70]

Human computers, mostly women, were vital in scientific calculations from the 18th century until the 1970s when electronic computers replaced them, leading to widespread job losses (Abbate, 2012). Women already faced wage disparities, with white female human computers earning 35–80% less than men and women of color facing even greater gaps (Pew Research Center, 2023). Workplace segregation and corporate policies, like marriage bars, further restricted their career growth (American Association of University Women, 2021).

Although many women transitioned into programming, discrimination persisted as men dominated the field when salaries rose (Card & DiNardo, 2002). Today, women in tech earn about 80–85% of men's wages due to systemic biases, workplace culture, and limited opportunities (Wilson and Darity, 2020). Efforts like diversity programs and pay transparency aim to close the gap, but historical inequalities continue to shape the industry (Center for American Progress, 2020).

Today, Artificial Intelligence continues this pattern, threatening to exacerbate existing inequalities. Prominent economists—including Amartya Sen, Joseph Stiglitz, Thomas Piketty, and Daron Acemoglu—argue that inequality is shaped by policy choices rather than inevitable economic forces. AI, as a human-made system, reflects these choices: it can either reinforce economic divides or be leveraged for broader social benefit.

Industry 5.0 takes the focus a step further and brings the human worker back into the equation, emphasising anthropocentric technology and be a game changer. The core idea is the

[70] https://ethw.org/Oral-History:Mary_Tsingou_Menzel#Advice_for_Women_in_Computing

collaboration between humans and advanced technologies like artificial intelligence, robotics, and automation. Unlike Industry 4.0, where automation often replaced human workers, Industry 5.0 seeks to create a synergy between humans and machines, allowing for customisation, creativity, and innovation[71].

However, key questions arise: Who benefits from AI-driven growth? Who bears the burdens of job displacement and economic disruption? Addressing these challenges requires intentional policies to ensure AI serves society equitably (Atkinson, 2025), rather than concentrating power and wealth among a privileged few. Inequality is not an immutable law of economics—Inequality is a choice. The path forward depends on how we choose to govern and distribute the benefits of AI. It should be noted that such choices are already being explored within the field of Computer Science, where concerned scholars are seeking to develop technologies that better reflect equitable distribution of resources on the basis of tools for cooperatives, development for social justice, localized development, (Sharma et al, 2023), and pro-labour design (Wolf & Dombrowski, 2022), in a direct response to the negative impacts of capitalism on societies.

3.9.3.1 Workforce Polarisation and AI-Driven Inequality

One of the most prominent ways in which AI contributes to inequality is through its differential impact on workers based on skill levels. Skilled workers, possessing the expertise necessary to effectively utilize AI tools, can significantly enhance their productivity, increasing their value and earning potential within the labour market (Gmyrek, 2023; Cazzaniga *et.al*., 2024). Conversely, unskilled workers are more vulnerable to automation, as their roles are increasingly replaced by AI-driven technologies. This trend limits their opportunities to adapt or remain competitive, thereby exacerbating income disparities both within and across firms.

The emergence of a two-tiered economy further underscores this divide. On one side are skilled professionals thriving in AI-enhanced roles, experiencing increased income and career advancement opportunities. On the other side are unskilled, younger, or older workers who face substantial challenges transitioning into roles demanding high levels of digital literacy, adaptability, and AI proficiency. Additionally, middle-skilled white-collar workers may also become casualties of AI-driven transformation, as their traditional roles are automated, forcing them either into unemployment or lower-paying service positions. This phenomenon, often referred to as the "hollowing out" of middle-tier roles, has significant implications for economic mobility, workforce polarisation, and broader societal inequality (Stiglitz, 2015, Georgieff, 2024; Gmyrek, 2023; Cazzaniga *et.al.*, 2024).

Workforce polarisation could potentially lead to decreased demand due to rising unemployment and low wages. Such conditions may trigger long-term economic stagnation despite technological advancements. Instead of greater prosperity, societies may experience economic hardship, as a reduced demand for labour could result in lower incomes and deteriorating living standards, even in scenarios where productivity and national income rise (Stiglitz, 2015).

3.9.3.2 Ownership and Capital

Unlike workers, managerial staff and capital owners are less directly affected by AI-driven automation. In fact, they are positioned to benefit significantly from the productivity enhancements and economic growth facilitated by AI. This asymmetry exacerbates the divide between labour and capital, intensifying structural inequalities within the economy (Moll, 2024; Tai, 2020; Cazzaniga *et.al.*, 2024). According to Yanis Varoufakis (2024), digital capitalism is giving rise to a novel

[71] https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en

economic order known as *technofeudalism,* wherein traditional concepts such as ownership and capital undergo profound transformation.

In *technofeudalism*, large technology platforms assume roles akin to feudal lords, deriving their power not primarily from ownership of traditional means of production but from control over essential digital infrastructures—algorithms, networks, and user-generated data. Under this system, ownership is no longer tied to physical assets or industrial capital but is instead embedded in monopolized digital platforms that regulate access, distribution, and participation in economic activity. Capital evolves into digital rent, extracted from users who supply labour, attention, and personal data. This process not only deepens inequality but also reinforces dependence on technology giants, which increasingly function as gatekeepers of social and economic interaction.

3.9.3.3 Corporate and Global Disparities

The disparities caused by Artificial Intelligence are not limited to individuals or workers; they also manifest at the organisational and global levels. Large companies, with their resources, expertise, and infrastructure, are well-positioned to adopt AI and harness its potential at scale. In contrast, small and medium-sized enterprises (SMEs) may struggle to compete, lacking the capital to implement AI effectively. Paradoxically, very small firms and freelancers might find their capacity for larger projects enhanced by AI, introducing new dynamics of competition. (Cazzaniga *et.al.*, 2024)

On a global scale, the digital divide becomes even more pronounced. Wealthier countries with robust digital infrastructure and research capabilities are better equipped to integrate AI into their economies, reaping its benefits. In contrast, poorer nations may face minimal disruption from AI but lack the means to harness it for growth. Even within regions like the European Union, disparities in AI adoption highlight the challenges faced by less affluent member states (Frankowska & Pawlik, 2022). These patterns underscore the potential for AI to exacerbate existing inequalities both within and between countries (Gmyrek, 2023; Cazzaniga *et.al.*, 2024).

3.9.3.4 AI and the Dignity of work

AI's influence on work and workplaces extends beyond automation and efficiency, often to the detriment of job quality, worker dignity, and the overall experience of the labour market. One overlooked issue is the increasing use of AI systems to monitor and evaluate employees, particularly in office environments. These systems track productivity by assessing work quality, measuring time spent on tasks, monitoring break frequency, and even analysing employees' social media activity (Riso and Litardi, 2024). The compiled data is often used to generate performance scores for supervisors to review, potentially creating a work culture centred on surveillance rather than trust. Whilst the AI Act explicitly bans AI-driven workplace surveillance, loopholes may allow employers to justify certain levels of monitoring under the guise of performance evaluation, or they may simply rely on weak enforcement mechanisms to continue these practices (Riso and Litardi, 2024). The result is a workplace where employees may feel pressured to conform to AI-driven metrics at the expense of well-being, creativity, and overall job satisfaction. Further, employers who are actively malicious and may look for ways to make the work lives of their employees more difficult are made far more capable of doing so through the use of AI monitoring software (Boddy & Ivory, 2024). The perception that AI systems are accurate can also enable abusive behaviour from management; for example, several hundred sub-postmasters of the UK Post Office wrongly faced criminal prosecution and some even took their own lives as malicious managers attributed the flaws of an accounting system (Horizon) to human error. This scandal demonstrated very effectively how someone malicious could use the cover of AI decisions or even manipulate them to hurt people, be it employees or others (Boddy & Ivory, 2024).

Beyond workplace surveillance, AI also reshapes the labour market through its role in hiring processes. A competitive "AI arms race" has emerged, where job seekers use AI to craft résumés and cover letters whilst employers deploy AI tools to sift through thousands of applications in

search of ideal candidates (Naveen, 2024). This mechanisation of hiring often leads to automated rejections of qualified applicants who might have otherwise excelled in a given role, entirely removing the human element from crucial hiring decisions. AI-driven hiring algorithms struggle to assess qualities that rely on social context and nuance, raising ethical concerns about job seekers' autonomy in self-definition and representation (Aizenberg, 2023). In contrast, human evaluation of applications ensures that candidates are considered as individuals rather than data points, preserving the dignity of applicants and making the hiring process a more authentic exchange of interest and opportunity.

3.10 AI in Healthcare

The integration of artificial intelligence (AI) into healthcare is accelerating, offering opportunities to enhance medical diagnosis, care personalisation, and precision. By harnessing its advanced pattern recognition capabilities, AI may assist physicians in making more accurate and timely decisions, ultimately improving patient outcomes and the quality of care. These advancements hold significant promise for reshaping healthcare delivery and addressing longstanding challenges in the field.

However, the adoption of AI also brings with it critical risks that must be addressed to ensure its safe and effective use. Concerns surrounding the governance of patient data, transparency in medical decision-making, and the preservation of patient trust and autonomy highlight the need for careful oversight. Without robust safeguards, these issues could undermine the integrity of healthcare systems, not only creating barriers to the adoption of AI, but also undermining public trust in the field. In order to fully realize the benefits AI can bring to healthcare whilst minimising unintended consequences, there needs to be an abundance of caution from policymakers to address these outstanding risks.

AI, mental health and emotion

One specific domain within which AI has seen recent development has been its use in support of mental health. According to Queensland Brain Institute (2023), one in two people will develop a mental health disorder in their lifetime, and more recent research into female brain health shows that all women in menopause will experience some alterations in their neurochemistry due to hormonal fluctuations (Mosconi, 2024), indicating that the proportion of those experiencing poor mental health is likely to be far higher. Whilst the global need for mental health support greatly outstrips the availability of robust interventions (World Health Organisation, 2020), the rush to meet demand with AI systems should be approached with caution. The use of AI in the mental health domain requires the processing, storage and analysis of highly sensitive personal and behavioural data and, whilst its use in clinical settings necessitates stringent regulation and oversight, the embedding of similar technology into companion systems such as chatbots, currently prevalent in the United States (Keierleber, 2022) provides a worrying loophole. The rise of chatbots and companion applications, particularly those that exhibit high level of anthropomorphism, create a new concerning design dynamic, particularly when used by those considered vulnerable.

This also speaks to a broader trend, within AI innovation, that has become known as Emotional AI. Systems such as these are designed to recognise, respond to and *influence* a user's emotional state, though not necessarily for therapeutic purposes and not always targeted at adults. A recent report, funded by the Office of the Privacy Commissioner of Canada (OPC) as part of the 2024-2025 Contributions Program outlined the prospective risks of AI systems designed to detect, stimulate or replicate emotional states in the context of children. Such systems tend to come in the form of toys or companions. Of particular concern is "when AI companions are marketed as 'friends' or for 'well-being' and imply that they are good for mental health without saying that outright, they fall outside of regulations that govern health or medical data" (Rosner & McStay, 2025, p.9). This legal loophole allows AI systems undue and unregulated access and influence over children's emotional state. Over longer-term use this also enables detection of indicators of protected characteristics such as sexuality, gender identity and other intimate data. The influential nature of such systems also means that such delicate processes in a child or adolescents' life might be

influenced, intentionally or otherwise. Rosner & McStay (2025) argue strongly that, in the context of children, the retention, analysis and further use of such data, even in the context of product improvement or marketing, should be banned, and that any development in this area requires robust guardrails. This is particularly the case where third parties, or brokers, are critical to the product's business model.

3.10.1 Diagnostics

Artificial intelligence (AI) has a long history in medicine, with its earliest recorded use dating back to 1976 when an algorithm was developed to identify causes of acute abdominal pain. Since then, AI has been deployed in a wide array of applications within healthcare, including disease detection and classification (Aung *et.al.*, 2021). For instance, AI has been employed in analysing medical images to identify skin cancers and retinopathy, classifying pathology in radiology scans, and delineating abnormalities in electrocardiograms. Beyond individual diagnostics, AI has also demonstrated utility in predicting disease patterns in epidemiology. During the COVID-19 pandemic, machine learning algorithms were employed to track the spread of the virus and identify potential outbreak clusters (Aung *et.al.,* 2021).

Whilst the deployment of AI in healthcare remains in its early stages, it holds vast potential to strengthen the field (Aung *et.al.*, 2021; Rashid & Kausik, 2024). One of its most promising applications lies in the analysis of large datasets to develop personalized treatment plans, advancing the scope of precision medicine (Elendu *et.al.*, 2023). As populations age and medical technology evolves, the complexity and volume of patient data have grown significantly. This increase, coupled with the prevalence of comorbidities, has complicated the work of physicians, who might find themselves unable to digest all the available information for each patient (Koski & Murphy, 2021). AI is being harnessed to reduce the workload of physicians; by aiding in early disease detection through the identification of patterns and anomalies in medical scans that might otherwise be overlooked. This capability not only improves patient outcomes but also reduces the financial burden on healthcare systems by minimising costly hospitalisations and enhancing preventative medicine efforts (Elendu *et.al.,* 2023; Koski & Murphy, 2021; Rashid & Kausik, 2024).

Prominent technology companies have partnered with healthcare institutions to deploy AI for disease detection and other applications. Examples include NVIDIA's Clara, which supports medical imaging and drug discovery at the National Institutes of Health and Massachusetts General Hospital; IBM Watson Health, which aids in diagnosis and drug discovery at the Mayo Clinic; and Zebra Medical Vision, which collaborates on medical imaging analysis with Stanford University Medical Center and Oxford University Hospitals NHS Foundation Trust (Rashid & Kausik, 2024). These collaborations highlight the increasing role of AI in clinical environments. Furthermore, AI is being leveraged to support clinical decision-making by assisting doctors with evidence-based treatment recommendations, analysing potential drug interactions, and improving diagnostic accuracy (Elendu *et.al.,* 2023). These tools empower physicians to make better-informed decisions and optimize patient care.

AI's role in high-level medical decision-making is noteworthy. Clinical Decision Support (CDS) systems have been in use since the 1970s and 1980s to reduce variations in care and improve adherence to medical guidelines (Koski & Murphy, 2021). Today, AI-enhanced Clinical Decision Support Systems (AI-CDSS) are employed in specialized settings, such as tumour board conferences. In these scenarios, AI analyses patient data alongside extensive databases of other cancer patients, simulating various treatment options and predicting their probabilities of success (Tretter *et.al*, 2023). Additionally, AI enables the creation of Digital Twins (DT) for specific organs, providing a virtual model that simulates interactions with medications and evaluates the potential impacts of medical interventions (Tretter, *et.al*., 2023).

The potential of AI to assist physicians in diagnosing and treating ailments is immense. By addressing the increasing workload faced by medical professionals and reducing the risks associated with human error, AI offers a pathway to more efficient and accurate healthcare

delivery. However, given the sensitive nature of healthcare, it is imperative to ensure the safe deployment of AI technologies.

3.10.2 The Role of AI in Services and Patient Care

AI is increasingly being explored and implemented as a versatile tool to assist in numerous aspects of healthcare beyond just diagnostics, branching into patient care, insurance processes, administrative tasks, and public health services. Similar to its applications in other professions, AI has the potential to automate certain repetitive administrative tasks that are often handled by physicians and nurses (Rashid & Kausik, 2024). These tasks, whilst essential, typically require minimal cognitive effort but can consume valuable time that could otherwise be dedicated to direct patient care. By offloading such duties to AI systems, healthcare professionals can focus more on the human elements of care (Aung *et.al.,* 2021).

For example, AI chatbots and virtual assistants are being employed to manage a range of patient interactions. They can efficiently handle routine queries, appointment scheduling, and even insurance verifications, significantly streamlining the patient experience (Elendu *et.al.,* 2023). Furthermore, AI has demonstrated its ability to process and manage large-scale question-based screenings, delivering results at a quality level comparable to medical staff (Aung *et.al.,* 2021). Automating these processes not only reduces the administrative burden on healthcare providers but also enhances overall efficiency, making healthcare systems more responsive to patient needs.

In addition to administrative support, AI has practical applications in assisting nurses with patient care. Remote monitoring systems powered by AI can track a patient's health from a distance, reducing the necessity for some in-person home visits. These technologies can alert nurses to potential issues in real time, enabling earlier interventions. AI can also assist nurses with documentation tasks and even provide virtual training and coaching, ensuring they remain equipped with the latest skills and knowledge (Koski & Murphy, 2021).

In countries like the United States, where private insurance dominates, AI-driven big data analysis is being utilized to reduce insurance costs, detect fraudulent claims, and personalize insurance services to better meet individual needs (Gehri, 2024; Ho *et.al.*, 2020). Such innovations not only enhance efficiency in private systems but also present opportunities to alleviate strain on universal healthcare systems in other countries by optimising resource allocation and reducing operational bottlenecks.

However, despite the numerous existing and anticipated benefits of AI in healthcare, these advancements come with significant risks that must be carefully managed. Ensuring that AI technologies are implemented responsibly and ethically is crucial if they are to truly improve lives.

3.10.3 Risks of AI in healthcare

The potential risks associated with the increased integration of AI systems into healthcare are significant and cannot be overlooked in the pursuit of greater efficiency or improved delivery of care. Policymakers must navigate a complex web of challenges, including patient data privacy, the possibility of introducing new biases, the erosion of trust in healthcare, concerns surrounding patient autonomy, and questions about the evolving responsibilities of physicians (Elendu *et.al.*, 2023, Bluemke *et.al*., 2023). These considerations are essential in determining whether the expansion of AI in healthcare is not only feasible but also beneficial in the long term. The decision to incorporate AI technologies should not be viewed as an inevitability, but rather as a deliberate choice that requires thorough evaluation and careful management to ensure that the risks do not outweigh the rewards.

Case study: The denial of Healthcare Lawsuit

In November 2023[72] a lawsuit emerged after an AI system used in healthcare erroneously denied a patient necessary care. The case highlights the complex interplay between technological innovation and ethical responsibility in high-stakes environments.

An AI algorithm, designed to expedite healthcare claims processing, made a decision that left the patient without access to vital treatment. The incident raised serious questions about the system's transparency. With the algorithm's "black box" approach, neither patients nor clinicians could understand the reasoning behind the denial, undermining trust in the technology. Central to the lawsuit were issues of bias and accountability. Critics argued that the algorithm might have been trained on historical data imbued with biases, potentially disadvantaging certain patient groups. The lack of an explainable decision-making process compounded these concerns, as it prevented affected individuals from effectively challenging the decision. Legal experts have pointed to negligence, noting that insufficient oversight and inadequate testing of the AI system may have breached the duty of care owed to patients. The case underscores a broader regulatory gap, calling for clear guidelines to ensure AI systems in healthcare meet ethical and legal standards. Stakeholders—including healthcare providers, insurance companies, AI developers, and regulators—are now under increasing pressure to adopt more transparent and accountable practices.

This incident serves as a crucial reminder that as AI technologies become more integrated into critical sectors like healthcare, robust ethical frameworks and regulatory oversight must be in place. Ensuring fairness, transparency, and accountability in AI systems is not only a technical challenge but also a moral imperative to protect patient rights and foster public trust.

3.10.3.1 Issues Regarding Data in healthcare

One of the most pressing concerns in the adoption of AI in healthcare lies in its reliance on vast amounts of patient data (Aung *et.al.,* 2021; Elendu *et.al.*, 2023; Ho *et.al.*, 2020). Machine learning algorithms, which are integral to AI diagnostic tools, require extensive datasets to function effectively. However, acquiring this data ethically is a challenge. It often involves securing the consent of thousands, if not millions, of patients—a process that can be both complex and prone to mishandling in the pursuit of progress. Medical institutions are understandably cautious about sharing such sensitive data, and when data is transferred between private companies for algorithm development, the risk of breaches grows exponentially (Aung *et.al.,* 2021).

Health data is among the most sensitive types of personal information, and its misuse or unauthorized access can have devastating consequences. A stark example of these risks occurred in 2018, when the NHS shared the data of 1.6 million patients with DeepMind, a Google subsidiary, without securing patient consent (Aung *et.al.,* 2021). This incident underscored the ethical and legal dilemmas that arise when handling sensitive medical information, especially when the data is used to train AI algorithms.

Beyond privacy concerns, the data itself may introduce further challenges. AI systems often rely on structured, quantifiable information, which can lead to the phenomenon of "data reduction." In this process, patients are effectively reduced to a collection of quantifiable metrics that a machine can process, potentially overlooking critical aspects of care such as emotional well-being, pain, and suffering—factors that require human interpretation (Tretter, *et.al.,* 2023). Additionally, datasets may lack generalizability, meaning they could fail to represent diverse populations

[72] https://www.forbes.com/sites/douglaslaney/2023/11/16/ai-ethics-essentials-lawsuit-over-ai-denial-of-healthcare/

accurately. This can result in biased or inaccurate model building, limiting the effectiveness of AI in real-world applications (Koski & Murphy, 2021).

Biases already embedded in historical health data present yet another problem. If these biases are reproduced in AI models, they could exacerbate existing disparities in treatment or insurance practices (Gehri, 2024). Worse still, AI systems, by their very nature, may lack the discretion to identify and address new biases introduced during their pattern recognition processes, further compounding the issue.

3.10.3.2 Issues with Accountability and Decision-Making

Another significant challenge in implementing AI in healthcare is the "black box" problem. Many AI systems, particularly those based on deep learning, operate in ways that are opaque to humans, making it difficult for clinicians to understand or verify the reasoning behind their recommendations (Aung *et.al.,* 2021; Gehri, 2024; Tretter, *et.al.,* 2023). This lack of transparency can understandably lead to discomfort among healthcare providers, who may be reluctant to trust or act on recommendations from systems they cannot fully comprehend (Koski & Murphy, 2021).

Accountability is a related concern. AI systems are not infallible, and in a field as high-stakes as healthcare—where decisions can mean the difference between life and death—the inability of AI to explain or take responsibility for its mistakes is a glaring limitation (Aung *et.al.,* 2021). If an AI system provides incorrect recommendations, who bears the responsibility? Physicians may find themselves in a precarious position, torn between trusting a system that is often statistically correct and relying on their own clinical judgment.

On the flip side, highly reliable AI systems could unintentionally diminish the autonomy of healthcare providers. Physicians may hesitate to countermand decisions made by AI, especially if they know that the system is statistically more accurate in most cases. This dynamic could undermine the human element of care and lead to a troubling overreliance on technology (Tretter, *et.al.,* 2023).

AI systems could also contribute to the already overwhelming stress faced by healthcare professionals. For example, integrating AI into diagnostic workflows may introduce an additional layer of "alert fatigue." Clinicians already manage a barrage of notifications, and adding AI-generated alerts could further increase stress in an already demanding work environment (Koski & Murphy, 2021).

3.10.3.3 AI and Trust in healthcare

One of the most crucial aspects of healthcare is the general societal trust in the integrity of the system itself. If the public lacks confidence in the success of their treatment, the competence and ethics of medical professionals, or the commitment of hospitals to their well-being, they may be hesitant to seek care or follow public health directives. This erosion of trust can lead to individuals avoiding necessary treatment or disregarding health guidelines, ultimately worsening societal health outcomes. A healthcare system relies not only on medical advancements but also on the belief that those advancements serve the best interests of the people.

The rise of AI in healthcare introduces significant risks to this trust. Studies on public perceptions of AI in healthcare suggest that whilst most individuals accept AI handling administrative tasks such as scheduling appointments and managing follow-ups (Witkowski *et.al*., 2024), concerns arise when AI becomes involved in direct patient care. These concerns include the fear of losing the human element in healthcare, issues surrounding patient privacy and autonomy, potential increases in costs, and the risk of data biases negatively affecting care quality (Witkowski *et.al.,* 2024; Richardson, 2021). If not addressed properly, these apprehensions could erode trust in AI-driven healthcare initiatives, reducing public willingness to embrace beneficial AI applications.

A decline in trust toward AI in healthcare could not only hinder the potential benefits of AI but also degrade healthcare outcomes overall. Ensuring trust in AI adoption requires prioritising ethical

considerations, fostering transparency, and actively involving the public in discussions about AI integration. Educating patients, respecting their concerns, and implementing policies that safeguard their interests are essential steps in maintaining trust. By taking a proactive approach—advancing regulations and addressing ethical issues before deployment—healthcare institutions can harness AI's benefits whilst preserving public confidence in the system (Witkowski *et.al.,* 2024; Richardson *et.al.*, 2021; Elendu *et.al.*, 2023).

3.11 Open-source as a Game Changer

Open-source AI has the potential to drive economic growth by enabling a broader range of businesses and individuals to tailor AI systems to their specific needs. By lowering entry barriers, open-source AI allows for greater participation in AI development, fostering innovation and collaborative development of AI technologies. Estimates of the economic value created by open-source software across the global economy can measure in the trillions of dollars as it is responsible for the upkeep of the largest coding languages and much of what makes the internet what it is today (Hoffmann *et.al.*, 2024). Open-source principles in AI could come to have similar positive economic effects. However, as AI becomes more widely accessible, the economic and employment impacts of automation and AI-driven efficiencies may also become more pronounced, necessitating careful regulatory and policy responses to mitigate unintended consequences.

The impact of open-source AI on AI democratisation is similarly significant. By making AI technology more accessible, smaller companies and individuals can develop and deploy their own models without reliance on large, resource-rich corporations. This decentralisation of AI innovation fosters greater competition and reduces the monopolistic control of a few dominant players. A notable example is China's Deepseek R1, an open-source model developed at a significantly lower cost than many competitors. Its accessibility allows businesses and individuals worldwide to build their own AI systems, harnessing powerful tools that were previously restricted to major AI corporations. This exemplifies the transformative potential of open-source AI in democratising access to advanced technologies (Edmond, 2025).

Despite these advantages, key challenges remain in ensuring that open-source AI truly levels the playing field. Whilst AI development may become more open, the acquisition of large, high-quality datasets remains largely within the financial reach of the most powerful companies, states, and investors. Additionally, compliance with regulations such as the EU's AI Act may impose financial burdens that smaller entities struggle to meet (Gobiet, 2023). This means that high-risk and high-impact AI systems could still be dominated by the largest p0layers unless the EU and EU state governments step in with oversight and subsidisation. Without such support, the development and deployment of riskier AI models may remain an exclusive domain, limiting the full democratising potential of open-source AI.

Transparency is another key advantage of open-source AI, as it allows for greater scrutiny of model design and functionality. Governments, regulators, and watchdog organisations can more easily identify biases, security risks, and ethical concerns that may otherwise be obscured by proprietary trade secrets. The widespread adoption of open-source principles in AI development could make it easier to enforce regulations like the AI Act, ensuring that AI systems meet ethical and safety standards. However, to maximize these benefits, the EU must ensure that compliance costs do not become prohibitive for smaller developers (Gobiet, 2023). If only the largest companies can afford to navigate complex regulatory requirements, the transparency benefits of open-source AI could be undermined, as riskier AI models would remain in the hands of a few dominant players.

Ultimately, whilst open-source AI offers significant benefits, it is unlikely to be a transformative "game changer" unless the balance between AI democratisation, regulatory compliance, and safety is carefully maintained. Policymakers must work to create an environment where AI innovation is both inclusive and responsible, ensuring that open-source principles contribute to

economic growth, democratic access, and transparency whilst addressing the challenges of regulation and risk management.

4. Policy Recommendations

These policy recommendations aim to address the social, economic, and ethical challenges associated with AI integration. AI governance requires clear definitions, strong ethical foundations, robust legal frameworks, and greater involvement from public institutions. The European Union should continue refining policies, guidelines, and strategic frameworks that help lawmakers craft regulations ensuring AI research, development, and deployment adhere to well-defined ethical and legal standards. The AI Act and GDPR must be continuously updated to keep pace with technological advancements, closing potential loopholes that could arise. The overarching goal is to enhance transparency and fairness in AI systems whilst establishing clear chains of accountability. Without such measures, there is a significant risk that responsibility for AI-related abuses and financial harm will be obfuscated behind complex corporate structures and opaque decision-making processes. Policymakers must ensure that accountability mechanisms are in place for both AI developers and end-users to address potential harms stemming from AI deployment.

A particularly critical area for AI governance is healthcare, where AI intersects with regulatory, ethical, and labour concerns. Whilst AI has demonstrated significant potential in medical applications, its deployment must be approached with caution. Policies should prioritize transparency in AI-driven healthcare decisions and ensure that these decisions are clearly communicated to the public. Additionally, data collection for AI applications in healthcare must adhere to strict, consent-based models to safeguard patient privacy and autonomy. Physician discretion should always take precedence over AI-generated diagnostics to maintain trust and integrity in medical decision-making. Well-regulated AI implementation will be essential to fostering public confidence in AI-assisted healthcare, whilst rushed or inadequately regulated deployment risks eroding trust in both AI-driven medicine and the broader healthcare system, with potentially severe consequences.

Protecting individual privacy in an era where AI can compile and analyse vast amounts of personal data requires stricter oversight of data brokers who exploit AI technologies in ways that violate GDPR and the AI Act. Enforcement of GDPR must be strengthened with stricter compliance measures and explicit informed consent at every stage of data collection, transfer, and sale. Social media corporations should not be permitted to collect and monetize user data under ambiguous consent agreements. By ensuring that only data brokers who obtain genuine informed consent can operate freely, stronger regulations can curb the widespread privacy violations that many users unknowingly experience.

The economic impact of AI is another pressing concern. Whilst AI-driven automation and productivity enhancements hold promise, they also pose challenges such as job displacement, labour force polarisation, and increasing inequality across industries and regions. Given the difficulty of predicting AI's full economic impact, policies must be flexible and continuously updated based on ongoing social research into AI's effects on labour markets. When AI leads to reduced work hours in certain professions, policies should prioritize redistributing work rather than resorting to layoffs. Where entire professions become obsolete due to AI, displaced workers must receive reskilling opportunities and financial support to maintain their living standards. Since AI is unlikely to generate new jobs at a rate sufficient to counteract job losses, proactive policies must facilitate worker transitions whilst minimising economic disruptions.

To equip workers with the skills needed to adapt to an AI-driven job market, AI training programs should be made widely accessible in relevant workplaces. Additionally, broader labour protections—such as workplace democracy, public ownership initiatives, stronger unionisation efforts, and even universal basic income—should be seriously considered. If the economic benefits

of AI remain concentrated among a small elite whilst its negative consequences disproportionately affect the most vulnerable, AI will exacerbate inequality rather than drive progress. To ensure AI benefits are equitably distributed across the EU, efforts should promote technology sharing, open-source development, and the fair allocation of AI expertise through targeted grants and funding initiatives.

AI's impact on workplace conditions must also be carefully managed. Employee dignity should be a primary consideration, extending beyond job security to include protections against AI-driven labour exploitation. Workplace surveillance, enabled by AI, presents a serious risk to workers' rights and mental well-being. The AI Act must be strictly enforced to prevent excessive monitoring, as employers have a financial incentive to circumvent regulations in pursuit of productivity gains.

The EU should also explore the creation of a union-wide hiring and job listing platform that limits the number of job applications individuals can submit and restricts the use of AI in candidate evaluations. Such a platform could transform hiring practices by reintroducing a human element into the process, reducing frustration and unfairness caused by AI-driven recruitment.

By implementing these comprehensive policy measures, AI research and deployment can be directed toward serving the broader public good rather than being dominated by corporate interests. AI should be harnessed as a tool for societal progress, ensuring its benefits are widely shared whilst minimising potential harms.

5. Conclusion

This review has mapped the rapidly evolving landscape of AI research, exposed critical gaps, and examined the multifaceted implications of AI across society, the economy, and governance. It underscores that, while AI holds transformative promise—from boosting productivity and advancing healthcare diagnostics to opening new frontiers in creativity and sustainability—these benefits will only materialize if we act deliberately to align innovation with Europe's core values of equity, transparency, and human dignity.

Key findings of this report reveal that public research must be strengthened through pan-European open-source platforms and interdisciplinary partnerships, particularly integrating social sciences and the humanities to anticipate ethical blind spots. Effective governance demands robust transparency mechanisms, enforceable rights around privacy and agency, and multistakeholder accountability frameworks that prevent AI from reinforcing existing power imbalances or eroding democratic norms. Equally urgent is the need for proactive social and labour policies—reskilling programmes, fair platform work regulations, and protections against AI-driven surveillance—to ensure that economic gains are shared broadly rather than concentrated among a privileged few.

Looking ahead, Europe has both the responsibility and the opportunity to lead a more inclusive AI agenda. By operationalising the policy recommendations detailed in Section 4—bolstering public R&D funding, enforcing the AI Act, fostering digital sovereignty, and championing ethical standards—Horizon Europe can catalyse innovation that serves the public good. Continued monitoring of AI's social impact, ongoing stakeholder engagement, and iterative regulation will be vital to adapt to emerging challenges. With decisive, values-driven action, Europe can model an AI ecosystem that not only drives scientific and economic progress but also safeguards human rights, strengthens social cohesion, and affirms the dignity of every citizen.

6. References

Abbate, J. (2012). *Recoding Gender: Women's Changing Participation in Computing.* MIT Press

Acemoglu, Daron, and James A. Robinson (2012). Why Nations Fail: The Origins of Power, Prosperity, and Poverty. New York: Crown Publishers

Agarwal, D., Naaman, M., & Vashistha, A. (2025). AI suggestions homogenize writing toward Western styles and diminish cultural nuances. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25) (pp. 1–21). ACM. https://doi.org/10.1145/3706598.3713564

AI Governance Alliance: Briefing Paper Series. (2024) World Economic Forum. Retrieved May 30, 2024 from https://www.weforum.org/publications/ai-governance-alliance-briefing-paper-series/

Aizenberg, E., Dennis,M.J., Hoven, van den J. (2023, Sept.,12). Examining the assumptions of AI hiring assessments and their impact on job seekers' autonomy over self-representation. In: *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01783-1

Ahmed, Nur; Wahed, Muntasir; Thompson, Neil C. (2023). *ARTIFICIAL INTELLIGENCE. The growing influence of industry in AI research. Industry is gaining control over the technology's future.* Retrieved January 29, 2025 from https://ide.mit.edu/wp-content/uploads/2023/03/0303PolicyForum_Ai_FF-2.pdf

American Association of University Women. (2021). *Systemic Racism and the Gender Pay Gap.* https://www.americanprogress.org/article/quick-facts-gender-wage-gap/

Askell, Amanda; Bai, Yuntao; Chen, Anna; Drain, Dawn; Ganguli, Deep; Henighan, Tom; Jones, Andy Joseph, Nicholas; Ben Mann, et al. (2021). A General Language Assistant as a Laboratory for Alignment. arXiv preprint arXiv:2112.00861 https://arxiv.org/abs/2112.00861

Atkinson, Anthony B. (2015). Inequality: What Can Be Done? Cambridge, MA: Harvard University Press, 2015

Aung, Y. Y. M., Wong, D. C. S., & Ting, D. S. W. (2021). The promise of artificial intelligence: A review of the opportunities and challenges of artificial intelligence in healthcare. British Medical Bulletin, 139(1), 4–15. https://doi.org/10.1093/bmb/ldab016

Bagenal, Jessamy et al. (2024) Generative AI: ensuring transparency and emphasising human intelligence and accountability. The Lancet, Volume 404, Issue 10468, 2142 – 2143

Balan, K., Jenni, S., Parsons, A., Gilberst, A., Collomosse, J. (2023). "EKILA: Synthetic Media Provenance and Attribution for Generative Art," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023, pp. 913-922, https://openaccess.thecvf.com/content/CVPR2023W/WMF/papers/Balan_EKILA_Synthetic_Media_Provenance_and_Attribution_for_Generative_Art_CVPRW_2023_paper.pdf

Birhane, A., Kasirzadeh, A., Leslie, D. et al. Science in the age of large language models.Nat Rev Phys 5, 277–280 (2023). https://doi.org/10.1038/s42254-023-00581-4

Bluemke, E., Collins, T., Garfinkel, B., Trask., A. (2023) Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases. http://arxiv.org/abs/2303.08956

Böhme, R., Köpsell, S. (2010). Trained to accept? a field experiment on consent dialogs. In proc. CHI '10, Atlanta, GA: ACM, 2403–2406

Boddy, C., & Ivory, C. (2024). The Future of Work: Artificial Intelligence, Ruthless Managers, Psychopathic Consequences. *ROBONOMICS: The Journal of the Automated Economy*, *5*, 66. Retrieved from https://journal.robonomics.science/index.php/rj/article/view/66

Buolamwini, J. (2024). Unmasking the bias in facial recognition algorithms. *MIT Sloan: Ideas Made to Matter*. Retrieved April 27, 2025, from https://mitsloan.mit.edu/ideas-made-to-matter/unmasking-bias-facial-recognition-algorithms

Cabanac, G., & Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. Journal of the Association for Information Science and Technology, 72(12), 1461–1476. https://doi.org/10.1002/asi.24495

Caporusso., N. (2023). Generative Artificial Intelligence and the Emergence of Creative Displacement Anxiety: Review. Res. Directs Psychol. Behav. 3, 1. https://doi.org/10.53520/rdpb2023.10795

Card, D., & DiNardo, J. E., (2002). *Technology and U.S. Wage Inequality: A Brief Look*. https://davidcard.berkeley.edu/papers/tech%20wage%20inequ.pdf?utm_source=chatgpt.com

Caro, M. (2025). In America's news deserts, Meta's retreat from fact-checking severs a last link to fact-based news. Poynter. 23rd January 2025. https://www.poynter.org/fact-checking/2025/effect-facebook-fact-checking-partnership-news-deserts/ (accessed 23.01.25)

Cazzaniga et.al. (2024). Gen-AI: Artificial Intelligence and the Future of Work. IMF Staff Discussion Note SDN2024/001, International Monetary Fund, Washington, DC.

CEDEFOP. (2020, October 7). Future employment growth | CEDEFOP. https://www.cedefop.europa.eu/en/tools/skills-intelligence/future-employment-growth

Chan, M. (2024). Analyzing the Impacts of A.I. on Employment and More in the Primary, Secondary, and Tertiary Sectors. Critical Debates in Humanities, Science and Global Justice, 2(1). https://criticaldebateshsgj.scholasticahq.com/article/91294-analyzing-the-impacts-of-a-i-on-employment-and-more-in-the-primary-secondary-and-tertiary-sectors

Chandhiramowuli, S., Taylor, A. S., Heitlinger, S. & Wang, D. (2024). Making Data Work Count. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), pp. 1- 26. doi: 10.1145/3637367

Chesterman S. (2024). Good models borrow, great models steal: intellectual property rights and generative AI. Policy Soc., puae006. https://doi.org/10.1093/polsoc/puae006

Ciftci, U. A., Yuksek, G., Demir, I. (2023). My face my choice: Privacy enhancing deepfakes for social media anonymisation, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1369-1379. https://openaccess.thecvf.com/content/WACV2023/papers/Ciftci_My_Face_My_Choice_Privacy_Enhancing_Deepfakes_for_Social_Media_WACV_2023_paper.pdf

Center for American Progress. (2020). *Quick Facts About the Gender Wage Gap*. https://www.americanprogress.org/article/playbook-for-the-advancement-of-women-in-the-economy/closing-the-gender-pay-gap/

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., Smith, N. A. (2021) *All That's "Human" Is Not Gold: Evaluating Human Evaluation of Generated Text*. http://arxiv.org/abs/2107.00061

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., de Oliveira, N. (2023) Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance, Patterns, Volume 4, Issue 10, 100857, ISSN 2666-3899, https://doi.org/10.1016/j.patter.2023.100857

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t

Criado-Perez, C. (2019). Invisible women: Exposing data bias in a world designed for men. Abrams Press.

Conroy, G.; Mallapaty, Sm. (2025). How China created AI model DeepSeek and shocked the world. Government policies, generous funding and a pipeline of AI graduates have helped Chinese firms create advanced LLMs. Nature, January 30, 2025. https://www.nature.com/articles/d41586-025-00259-0?utm_source=chatgpt.com

Dathathri, S., See, A., Ghaisas, S. et al., (2024). Scalable watermarking for identifying large language model outputs. Nature 634, 818–823. https://doi.org/10.1038/s41586-024-08025-4

Davenport, C., (2025). This New Open-Source Alternative to Google Docs and Notion Is Backed by France and Germany. How-to-Geek. https://www.howtogeek.com/docs-alternative-google-docs-notion-france-germany/

De Cremer, D., & Kasparov, G. (2022). The ethics of technology innovation: A double-edged sword? AI and Ethics, 2(4), 533–537. https://doi.org/10.1007/s43681-021-00103-x

Della Croce, Y. Epistemic Injustice and Nonmaleficence. Bioethical Inquiry 20, 447–456 (2023). https://doi.org/10.1007/s11673-023-10273-4

DeepSeek LLM: Xiao Bi et al. 2024. Scaling Open-Source Language Models with Longtermism. https://arxiv.org/abs/2401.02954

Diakopoulos, N., (2019). Automating the News: How Algorithms Are Rewriting the Media. Cambridge, Massachusetts: Harvard University Press.

Ding, J., Akiki, C., Jernite, Y., Steele, A. L., Popo. T., (2023). Towards Openness Beyond Open Access: User Journeys through 3 Open AI Collaboratives. http://arxiv.org/abs/2301.08488

Diresta, R & Goldstein, J. (2024). How spammers and scammers leverage AI-generated images on Facebook for audience growth. Misinformation Review. Harvard Kennedy School. https://misinforeview.hks.harvard.edu/article/how-spammers-and-scammers-leverage-ai-generated-images-on-facebook-for-audience-growth/

Doo, F. X., Naranjo, W. G., Kapouranis, T., Thor, M., Chao, M., Yang, X., & Marshall, D. C. (2025). Sex-Based Bias in Artificial Intelligence-Based Segmentation Models in Clinical Oncology. Clinical Oncology (Royal College of Radiologists), 39, 103758. https://doi.org/10.1016/j.clon.2025.103758

Edmond, C. (2025, February 5). What is open-source AI and how could DeepSeek change the industry? World Economic Forum. Retrieved from https://www.weforum.org/stories/2025/02/open-source-ai-innovation-deepseek/

Edwards, B. (2025) AI search engines cite incorrect news sources at an alarming 60% rate, study says. Arstechnica. https://arstechnica.com/ai/2025/03/ai-search-engines-give-incorrect-answers-at-an-alarming-60-rate-study-says/

Elendu, C., Amaechi, D. C., Elendu, T. C., Jingwa, K. A., Okoye, O. K., John Okah, M., Ladele, J. A., Farah, A. H., & Alimi, H. A. (2023). Ethical implications of AI and robotics in healthcare: A review. Medicine, 102(50), e36671. https://doi.org/10.1097/MD.0000000000036671

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An Early Look at the LabourMarket Impact Potential of Large Language Models (No. arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: Picador, St. Martin's Press

European Commission. (2009-2020). AI Watch. AI Landscape Dashboard. https://ai-watch.ec.europa.eu/tools/ai-landscape-dashboard_en

Eurostat. (2024). Businesses in the manufacturing sector. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Businesses_in_the_manufacturing_sector

Frankowska, A., & Pawlik, B. (2022). A Decade of Artificial Intelligence Research in the European Union: A Bibliometric Analysis. In C. Biele, J. Kacprzyk, W. Kopeć, J. W. Owsiński, A. Romanowski, & M. Sikorski (Eds.), Digital Interaction and Machine Intelligence (pp. 52–62). Springer International Publishing. https://doi.org/10.1007/978-3-031-11432-8_5

Fukuyama, F. (2016). Governance: What Do We Know, and How Do We Know It? Annual Review of Political Science, 19(1), 89–105. https://doi.org/10.1146/annurev-polisci-042214-044240

Gartry, L. (2024). Protecting Public Interest Journalism while Personalising the News. Journalist Fellowship Paper. Reuters Institute https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-07/RISJ%20Journalist%20Fellowship%20Paper_Laura%20Gartry_TT24_Review.pdf

Gehri, J. (2024). AI Ethics in Insurance J.Gehri 25.11. Balzac Publishers. https://www.academia.edu/125857852/AI_Ethics_in_Insurance_J_Gehri_25_11

Georgieff. (2024, April 9). Artificial intelligence and wage inequality. OECD. https://www.oecd.org/en/publications/artificial-intelligence-and-wage-inequality_bf98a45c-en.html

Gmyrek, P., Berg, J., Bescond, D., & International Labour Organisation. Research Department. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality. ILO. https://doi.org/10.54394/FHEM8239

Gobiet, M. (2023, September 27). Transparency and Innovation: The Future of Open-Source AI. Onlim. Retrieved April 27, 2025, from https://onlim.com/en/transparency-and-innovation-the-future-of-open-source-ai/

Goktas, P. (2024). Ethics, transparency, and explainability in generative ai decision-making systems: a comprehensive bibliometric study. Journal of Decision Systems, 1–29. https://doi.org/10.1080/12460125.2024.2410042

Graylin, A. W., & Triolo, P. (2025). There can be no winners in a US-China AI arms race. MIT Review (21.01.25) https://www.technologyreview.com/2025/01/21/1110269/there-can-be-no-winners-in-a-us-china-ai-arms-race/

Grierson, J. (2023). Photographer admits prize-winning image was AI-generated. Guardian online (17th April 2023) https://www.theguardian.com/technology/2023/apr/17/photographer-admits-prize-winning-image-was-ai-generated

Haider, J., Söderström, K. R., Ekström, B., & Rödl, M. (2024). GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-156

Hall, P., Elllis, D. (2023). "A systematic review of socio-technical gender bias in AI algorithms", Online Information Review, Vol. 47 No. 7, pp. 1264-1279. https://doi.org/10.1108/OIR-08-2021-0452

Hall, R., Wilmot, C. (2025) Meta faces Ghana lawsuits over impact of extreme content on moderators. Guardian online https://www.theguardian.com/technology/2025/apr/27/meta-faces-ghana-lawsuits-over-impact-of-extreme-content-on-moderators

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., & Tang, H. (2024). Synthetic Data in AI: Challenges, Applications, and Ethical Implications. arXiv preprint arXiv:2401.01629.

Ho, C. W. L., Ali, J., & Caals, K., (2020). Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. Bulletin of the World Health Organisation, 98(4), 263–269. https://doi.org/10.2471/BLT.19.234732

Ho, S., Burke, G., (2022). An algorithm that screens for child neglect raises concerns. AP. https://apnews.com/article/child-welfare-algorithm-investigation-9497ee937e0053ad4144a86c68241ef1

Hoffmann, M., Nagle, F., & Zhou, Y. (2024). The Value of Open Source Software (Harvard Business School Strategy Unit Working Paper No. 24-038). SSRN. https://doi.org/10.2139/ssrn.4693148

Hoffman-Andrews, J. (2024). AI Watermarking Won't Curb Disinformation. Electronic Frontier Foundation (5th January 2024) https://www.eff.org/deeplinks/2024/01/ai-watermarking-wont-curb-disinformation

IAPP; Casovan, A., Jones, J., & Chaudhry, U. (2024). AI Governance in Practice Report 2024. International Association of Privacy Professionals. Retrieved April 27, 2025, from https://iapp.org/resources/article/ai-governance-in-practice-report/

IEEE Spectrum; Strickland, Eliza (2024). 15 Graphs That Explain the State of AI in 2024. The AI Index tracks the generative AI boom, model costs, and responsible AI. IEEE Spectrum, April 15, 2024. https://spectrum.ieee.org/ai-index-2024

Jiang, L. & Goetz, S. (2024). Artificial Intelligence Exploring the Patent Field. https://arxiv.org/pdf/2403.04105v1

Johnson, B., Bartola, J., Angell, R., Witty, S., Giguere, S., Brun, Y. (2024) Fairkit, fairkit, on the wall, who's the fairest of them all? Supporting fairness-related decision-making, EURO Journal on Decision Processes, Vol. 11, 100031, ISSN 2193-9438, https://doi.org/10.1016/j.ejdp.2023.100031.

Jones, B., Jones R., & Luger, E. (2022). AI 'Everywhere and Nowhere': Addressing the AI Intelligibility Problem in Public Service Journalism, Digital Journalism, 10:10, 1731-1755, https://doi.org/10.1080/21670811.2022.2145328

Jones, N., (2024). The AI revolution is running out of data. What can researchers do? 11th Dec 20204 Nature. https://www.nature.com/articles/d41586-024-03990-2

Kakhiani, D., (2024). Code at the Speed of Thought: Exploring the Impact of AI on Coding Practices. ResearchGate. https://doi.org/10.6084/m9.figshare.25664439.v1

Karippacheril, (2024). What we're reading about the age of AI, jobs, and inequality. World Bank Blogs. https://blogs.worldbank.org/en/jobs/What-we-re-reading-about-the-age-of-AI-jobs-and-inequality

Kaveh, A & Eisenberg, D. (2023). "Shaping the Future of Work: Responsible Design and Public Policy for Generative AI" (2023). AMCIS 2023 TREOs. 123. https://aisel.aisnet.org/treos_amcis2023/123

Keierleber, M. (2022). Young and depressed? Try Woebot! The rise of mental health chatbots in the US. The Guardian https://www.theguardian.com/us-news/2022/apr/13/chatbots-robot-therapists-youth-mental-health-crisis

King, J., Meinhardt, C. (2024). Rethinking Privacy in the AI Era Policy Provocations for a Data-Centric World. White Paper. Centre for Human-Centered Artificial Intelligence, Stanford University. https://hai.stanford.edu/sites/default/files/2024-02/White-Paper-Rethinking-Privacy-AI-Era.pdf

Kingsley, S., Sinha, P., Wang, C., Eslami, M., Hong, J. I., (2022). "Give Everybody [..] a Little Bit More Equity": Content Creator Perspectives and Responses to the Algorithmic Demonetisation of Content Associated with Disadvantaged Groups. Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 424 (November 2022), 37 pages. https://doi.org/10.1145/3555149

Knibbs, K. (2024). AI Slop Is Flooding Medium: The blogging platform Medium is facing an influx of AI-generated content. CEO Tony Stubblebine says it "doesn't matter" as long as nobody reads it. Wired, 28.10.2024

Koski, E., & Murphy, J. (2021). AI in Healthcare. In Nurses and Midwives in the Digital Age (pp. 295–299). IOS Press. https://doi.org/10.3233/SHTI210726

Kreps, S. E., McCain, R. M., & Brundage, M. (2020). All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation [Preprint]. SSRN. https://doi.org/10.2139/ssrn.3525002

Kurtzig, A (2025) The AI lie: how trillion-dollar hype is killing humanity. TechRadar (25th January 2025) https://www.techradar.com/pro/the-ai-lie-how-trillion-dollar-hype-is-killing-humanity

Lee, K., Cooper, A. F., Grimmelmann, J., and Grimmelmann, Ippolito, J., Ippolito, D. (2023). AI and Law: The Next Generation. Available at SSRN: https://ssrn.com/abstract=4580739 or http://dx.doi.org/10.2139/ssrn.4580739

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan,A., See, D., Noothigattu, R., Lee, S., Psomas, A., Procaccia, A. D. (2019). WeBuildAI: Participatory framework for algorithmic governance. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 181 (Nov. 2019), 35 pages. https://doi.org/10.1145/3359283

Lee, P. (2016). Learning from Tay's introduction. Official Microsoft Blog, Available at: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (accessed 10.01.25)

Lessig, L. (2006). Code and Other Laws of Cyberspace, Version 2.0. New York: Basic Books

Loewen, P. J., Lee-Whiting, B., Arai, M., Bergeron, T., Galipeau, T., Gazendam, I., Needham, H., Slinger, L., Yusypovych, S., (2024). Global Public Opinion on Artificial Intelligence (GPO-AI). Schwartz Reisman Institute for Technology and Society.https://srinstitute.utoronto.ca/public-opinion-ai

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

Longpre, S. et.al. (2024). *Bridging the Data Provenance Gap Across Text, Speech and Video*. Artificial Intelligence; Computation and Language; Computers and Society; Machine Learning; Multimedia Cites:arXiv:2412.17847 https://doi.org/10.48550/arXiv.2412.17847Focus to learn more

Li, J., Cao, H., Lin, L., Hou, Y., Zhu, R., El Ali, A. (2024). User Experience Design Professionals' Perceptions of Generative Artificial Intelligence. In Proc. CHI '24. ACM, New York, NY, USA, Article 381, 1–18. https://doi.org/10.1145/3613904.3642114

Lomas, N. (2025). European tech industry coalition calls for 'radical action' on digital sovereignty — starting with buying local. TechCrunch. https://techcrunch.com/2025/03/16/european-tech-industry-coalition-calls-for-radical-action-on-digital-sovereignty-starting-with-buying-local/

Lovato, J., Zimmerman, J. W., Smith, I., Dodds, P., & Karson, J. L. (2024). Foregrounding Artist Opinions: A Survey Study on Transparency, Ownership, and Fairness in AI Generative Art. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7(1), 905-916. https://doi.org/10.1609/aies.v7i1.31691

Luger, E. (2023). What do we know and what should we do about AI? Sage

Luger, E., Rodden, T. (2013). An Informed View on Consent for Ubicomp. In Proc. Ubicomp'13. ACM. https://www.researchgate.net/profile/Ewa-Luger-3/publication/266654104_An_informed_view_on_consent_for_UbiComp/links/56582f0d08ae4988a7b6c8e0/An-informed-view-on-consent-for-UbiComp.pdf

Marcin, S. (2019). Economic impacts of artificial intelligence. EPRS.

Mardiani E. and Iswahyudi, M.S. (2023). Mapping the Landscape of Artificial Intelligence Research: A Bibliometric Approach. In: West Science Interdisciplinary Studies 1(08):587-599. DOI:10.58812/wsis.v1i08.183

Maxwell, T. (2025) Microsoft's Satya Nadella Pumps the Brakes on AI Hype. Gizmodo. https://gizmodo.com/microsofts-satya-nadella-pumps-the-breaks-on-ai-hype-2000566483

Meagher, A., & Robertson, B. (2024). *Title of article*. *Journal Name, 12*(3), 123–145. https://doi.org/xx.xxx/yyyy

Metz, R. (2022). AI won an art contest, and artists are furious. CNN (3rd September 2022). https://edition.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html

Meyer, D. (2025). U.K. drops AI safety focus and signs up Anthropic to help transform public services. Fortune. https://fortune.com/2025/02/13/uk-ai-security-institute-safety-anthropic-trump-vance/

Milne, G. (2020). Smoke and Mirrors: How Hype Obscures the Future and How to See Past It. Robinson Press

Miziołek, T. (2021). Employing artificial intelligence in investment management. In The Digitalisation of Financial Markets. Routledge.

Moll. (2024). Uneven Growth: Automation's Impact on Income and Wealth Inequality | Request PDF. ResearchGate. https://doi.org/10.3982/ECTA19417

Monday &. Strappelli, (2024). BBC Does provenance build trust? Researching the impacts of showing media provenance information to audiences https://www.bbc.co.uk/rdnewslabs/news/does-provenance-build-trust/ (accessed 17.01.25)

Mosconi, L. (2024). The Menopause Brain. Allen & Unwin, c/o Atlantic Books, London

Mukherjee, S. (2024, May 28). Klarna using GenAI to cut marketing costs by $10 mln annually. Reuters. https://www.reuters.com/technology/klarna-using-genai-cut-marketing-costs-by-10-mln-annually-2024-05-28/

Naveen, Palanichamy (2024). The rise of AI in job applications: a generative adversarial tug-of-war. In: AI & SOCIETY https://doi.org/10.1007/s00146-024-02054-3

Newfield C. (2025). Humanities Decline in Darkness: How Humanities Research Funding Works. Public Humanities. 2025;1:e31. doi:10.1017/pub.2024.39

Newman-Griffis, D. (2025). AI Thinking: a framework for rethinking artificial intelligence in practice. R. Soc. Open Sci.12241482. http://doi.org/10.1098/rsos.241482

Nguyen, T. (2024). California governor signs laws to crack down on election deepfakes created by AI. Associated Press (AP) News, 18th September 2024, https://apnews.com/article/california-artificial-intelligence-deepfakes-election-0e70cb32b06d9187eaef5bdacaba6d77 (accessed 04.01.25)

Nigatu, H. H., Tonja, A. L., Rosman, B. Solorio, T., Choudhury, M. (2024) In proc. 2024 Conference on Empirical Methods in Natural Language Processing, 17753–17774 November 12-16, 2024. Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.983.pdf

Nizhnichenkov, SV., et al. (2023). Explaining Knock-on Effects of Bias Mitigation. In: *Machine Learning (cs.LG)*; Computers and Society (cs.CY). http://arxiv.org/abs/2312.00765

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: New York University Press.

Nockels, J., Gooding, P., Ames, S. et al. (2022) Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. Arch Sci 22, 367–392. https://doi.org/10.1007/s10502-022-09397-0

Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. AI & Society, 39, 1871–1882. https://doi.org/10.1007/s00146-023-01635-y

O'Donnell, J., Heaven, W. D., Heikkilä, M. (2025). What's Next for AI in 2025. MIT Technology Review (8th January 2025). https://www.technologyreview.com/2025/01/08/1109188/whats-next-for-ai-in-2025/

OPC; Internet of Things Privacy Forum. (2025). The Machine-Readable Child: Governance of Emotional AI used with Canadian Children (Contribution No. CP-000073). Office of the Privacy Commissioner of Canada. Retrieved April 27, 2025, from https://search.open.canada.ca/grants/record/opc-cpvp,CP-000073,current

Oduro, S. and Kneese, T. (2024). AI Governance needs Sociotechnical Expertise. Why the Humanities and Social Sciences are Critical to Government Efforts. Data & Society https://datasociety.net/wp-content/uploads/2024/05/DS_AI_Governance_Policy_Brief.pdf

Ozimek, P., Lainas, S., Bierhoff, HW. et al. (2023) How photo editing in social media shapes self-perceived attractiveness and self-esteem via self-objectification and physical appearance comparisons. BMC Psychol 11, 99. https://doi.org/10.1186/s40359-023-01143-0

Pew Research Center (2023). The Enduring Grip of the Gender Pay Gap. https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/

Piketti, Thomas (2014). Capital in the Twenty-First Century. Cambridge, MA: Belknap Press of Harvard University

Piketti, Thomas (2020). Capital and Ideology. Belknap Press

Queensland Brain Institute. (2023). Half of World's Population Will Experience a Mental Health Disorder. Harvard Medical School, News & Research. https://hms.harvard.edu/news/half-worlds-population-will-experience-mental-health-disorder

Rahman-Jones, I. (2025). Man files complaint after ChatGPT said he killed his children. BBC News. https://www.bbc.co.uk/news/articles/c0kgydkr516o

Rashid, A. B., & Kausik, M. A. K. (2024). AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. Hybrid Advances, 7, 100277. https://doi.org/10.1016/j.hybadv.2024.100277

Rauktis, M., McCrae, J. (2010). The Role of Race in Child Welfare System Involvement in Allegheny County. Allegheny County DHS. https://www.alleghenycountyanalytics.us/wp-content/uploads/2015/12/The-Role-of-Race-in-Child-Welfare-System-Involvement-in-Allegheny-County.pdf

Regona, M., Yigitcanlar, T., Xia, B., & Li, R. Y. M. (2022). Opportunities and Adoption Challenges of AI in the Construction Industry: A PRISMA Review. ResearchGate. https://doi.org/10.3390/joitmc8010045

Ren, J., Xu, H., He, P., Cui, Y., Zeng, S., Zhang, J., Wen, H., Ding, J., Liu, H., Chang, Y., Tang, J. (2024). Copyright Protection in Generative AI: A Technical Perspective. http://arxiv.org/abs/2402.02333

Rezk, A. M., Simkute, A., Luger, E., Vines, J., Elsden, C., Evans, M., Jones R. (2024). Agency Aspirations: Understanding Users' Preferences And Perceptions Of Their Role In Personalised News

Curation. In Proc. Conference on Human Factors in Computing Systems (CHI '24). ACM, New York, NY, USA, Article 190, 1–16. https://doi.org/10.1145/3613904.3642634

Riso, S. and Litardi, Ch. (2024). Employee monitoring: A moving target for regulation, in: Employment and labour markets. Eurofund. https://www.eurofound.europa.eu/en/resources/article/2024/employee-monitoring-moving-target-regulation

Richardson, J.P., Smith, C., Curtis, S., Watson, S, Zhu, X., Barry, B., and Sharp, R.R. (2021). Patient apprehensions about the use of artificial intelligence in healthcare. In: Digital Medicine https://doi.org/10.1038/s41746-021-00509-1

Romano, A. (2019). A group of YouTubers is trying to prove the site systematically demonetizes queer content. Vox. https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetisation-nerd-city-algorithm-report

Rosner, G., McStay, A. (2025). The Machine-Readable Child: Governance of Emotional AI Used with Canadian Children. (forthcoming)

Runciman, D. (2023). The Handover: How We Gave Control of Our Lives to Corporations, States and Ais. Profile Books

Ruschemeier, H. (2023). Data Brokers and European Digital Legislation. European Data Protection Law Review. Jahrgang 9, Ausgabe 1 (2023), pp. 27 – 38. DOI: https://doi.org/10.21552/edpl/2023/1/7

Russell, M., Renwick, A., James, L. (2022). What is Democratic Backsliding, and is the UK at Risk? The Constitutional Unit, Briefing. UCL, https://www.ucl.ac.uk/constitution-unit/sites/constitution_unit/files/backsliding_-_final_1.pdf

Sanderson, C., Schleiger, E., Douglas, D., Kuhnert, P., & Lu, Q. (2024). Resolving Ethics Trade-offs in Implementing Responsible AI [Preprint]. arXiv.

Sarridis, Ioannis et al. (2023). TOWARDS FAIR FACE VERIFICATION: AN IN-DEPTH ANALYSIS OF DEMOGRAPHIC BIASES. https://arxiv.org/pdf/2307.10011

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's Next for AI Ethics, Policy, and Governance? A Global Overview. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 153–158). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375804

Schjøtt Hansen, A., & Hartley, J. M. (2021). Designing What's News: An Ethnography of a Personalisation Algorithm and the Data-Driven (Re)Assembling of the News. Digital Journalism, 11(6), 924–942. https://doi.org/10.1080/21670811.2021.1988861

Sen, Amartya, K. (1995). Inequality Reexamined. Cambrige MA: Harvard University Press.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B. Y. (2023). Glaze: protecting artists from style mimicry by text-to-image models. In Proc 32nd USENIX Conference on Security Symposium (SEC '23). USENIX Association, USA, Article 123, 2187–2204

Sharma, V., Kumar, N., Nardi, B. (2023) Post-growth Human–Computer Interaction. ACM Trans. Comput.-Hum. Interact. 31, 1, Article 9 (November 2023), 37 pages. https://doi.org/10.1145/3624981

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., and Zhao, J. (2023). Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. In '2023 International Conference on Computer and Applications (ICCA)', November 2023. 1–7. https://doi.org/10.1109/ICCA59364.2023.10401723

Simkute, A., Luger, E., Evans, M., Jones, R. (2024). "It is there, and you need it, so why do you not use it?" Achieving better adoption of AI systems by domain experts, in the case study of natural science research. https://doi.org/10.48550/arXiv.2403.16895

Simon, F. M. (2022). Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy. Digital Journalism, 10(10), 1832–1854. https://doi.org/10.1080/21670811.2022.2063150

Spencer, D. A. (2024). AI, automation and the lightening of work. AI & SOCIETY. https://doi.org/10.1007/s00146-024-01959-3

Sterling, T. (2025). Dutch parliament calls for end to dependence on US software companies. Reuters. https://www.reuters.com/world/europe/dutch-parliament-calls-end-reliance-us-software-2025-03-18/

Strauß, N., Huber, B., & Gil de Zúñiga, H. (2021). Structural Influences on the News Finds Me Perception: Why People Believe They Don't Have to Actively Seek News Anymore. Social Media + Society, 7(2). https://doi.org/10.1177/20563051211024966

Stiglitz, Joseph (2012). The Price of Inequality: How Today's Divided Society Endangers Our Future. New York: W.W. Norton & Company

Stiglitz, Joseph (2013). Inequality is a choice. The New York Times. https://www3.nd.edu/~pweithma/Readings/Cohen,%20Gerald/Stiglitz%20(Inequality%20Is%20a%20Choice%20-%20NYTimes).pdf

Stiglitz, Joseph (2015). The Great Divide: Unequal Societies and What We Can Do About Them. W.W. Norton & Company.

Stiglitz Joseph (2015). Joseph Stiglitz on Inequality and Economic Growth. Ford Foundation. https://www.fordfoundation.org/news-and-stories/big-ideas/inequalityis/joseph-stiglitz-on-inequality-and-economic-growth/

Stiglitz, Joseph (2019). People, Power and Profits. Progressive Capitalism for an Age of Discontent. W.W. Norton & Company

Sudalairaj, S., Bhandwaldar, A., Pareja, A., Xu, K., Cox, D. D., & Srivastava, A. (2024). LAB: Large-Scale Alignment for ChatBots. arXiv preprint arXiv:2403.01081.

Swenson, A., & Chan, K. (2024). Election disinformation takes a big leap with AI being used to deceive worldwide. Associated Press (AP) News, 14th March 2024, https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd (accessed 06.01.25)

Tai, M. C.-T. (2020). The impact of artificial intelligence on human society and bioethics. Tzu-Chi Medical Journal, 32(4), 339–343. https://doi.org/10.4103/tcmj.tcmj_71_20

Terras M, Anzinger B, Gooding P et al. (2025) The artificial intelligence cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for automated text recognition [version 1; peer review: 1 approved with reservations, 1 not approved]. Open Res Europe 2025, **5**:16 (https://doi.org/10.12688/openreseurope.18747.1)

Thieme, A., Nori, A., Ghassemi, M., Bommasani, R., Osman Andersen,T., Luger, E. (2023) Foundation Models in Healthcare: Opportunities, Risks & Strategies Forward. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 512, 1–4. https://doi.org/10.1145/3544549.3583177

Tollon, F., & Vallor, S. (2025). The Responsible AI Ecosystem: A BRAID Landscape Study. (forthcoming)

Tretter, M., Ott, T., & Dabrock, P. (2023). AI-produced certainties in health care: Current and future challenges. AI and Ethics. https://doi.org/10.1007/s43681-023-00374-6

Tyson, L. D., & Zysman, J. (2022). Automation, AI & Work. Daedalus, 151(2), 256–271.

UNFAO. (2022). Labour impacts of agricultural automation. https://doi.org/10.4060/cb9479en

United Nations Environment Programme (2024) Story: 'AI has an environmental problem. Here's what the world can do about that'. https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about (accessed 03.01.25)

Varoufakis, Yanis (2023). Technofeudalism. What killed Capitalism. UK. Penguin

Vincenzi, B., Stumpf, S., Taylor, A. S., Nakao, Y. (2024). Lay User Involvement in Developing Human-Centric Responsible AI Systems: When and How? ACM J. Responsible Comput., 3652592. https://doi.org/10.1145/3652592

Wachter, R. M., & Brynjolfsson, E. (2024). Will Generative Artificial Intelligence Deliver on Its Promise in Health Care? JAMA 331, 1 (January 2024), 65–69. https://doi.org/10.1001/jama.2023.25054

Wagner, B. (2018). Ethics as an escape from regulation: From “ethics-washing” to ethics-shopping? In E. Bayamlioglu, I. Baraliuc, L. A. W. Janssens, & M. Hildebrandt (Eds.), Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen (pp. 84–89). Amsterdam University Press. https://doi.org/10.1515/9789048550180-016

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. Available at https://ainowinstitute.org/discriminating-systems

Wilkinson, I. (2025). Trump, Stargate, DeepSeek: A new, more unpredictable era for AI? Chatham House. https://www.chathamhouse.org/2025/02/trump-stargate-deepseek-new-more-unpredictable-era-ai

Wilson, V. and Darity, W. Jr. (2022). Understanding black-white disparities in labourmarket outcomes requires models that account for persistent discrimination and unequal bargaining power. In: Economic Policy Institute. https://www.epi.org/unequalpower/publications/understanding-black-white-disparities-in-labor-market-outcomes/

Wirtz, J., & Pitardi, V. (2023). How intelligent automation, service robots, and AI will reshape service products and their delivery. Italian Journal of Marketing, 2023(3), 289–300. https://doi.org/10.1007/s43039-023-00076-1

Witkowski, K., Okhai, R., Neely, S. R. (2024). Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. In: BMC Medical Ethics. https://doi.org/10.1186/s12910-024-01066-4

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., & West, S. M. (2019). AI Now 2019 Report: The Social and Economic Implications of Artificial Intelligence for Law and Policy. AI Now Institute, New York University.

World Health Organisation (2020). Mental Health Atlas. WHO https://www.who.int/publications/i/item/9789240036703

Wolf, C.T., Asad, M., and Dombrowski, L.S. (2022). Designing within capitalism. Proc. Of the 2022 ACM Designing Interactive Systems Conference. ACM, 2022, 439–453.

Wright, K., Scott, M., Bunce, M. (2024). Capturing News, Capturing Democracy: Trump and the Voice of America. OUP USA

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union.
You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Artificial Intelligence is transforming Europe—delivering breakthroughs in precision medicine and climate resilience—while the opacity of “black-box” systems and the spread of misinformation threaten public trust and risk widening inequalities. This Commission-mandated independent expert review reveals deep divides between well-funded private-sector innovation and under-resourced public research, alongside stark geographic imbalances among Member States and overlooked potential for AI in social-good and environmental initiatives. To realign AI with European values of equity, transparency and human dignity, it calls for significantly increased public R&D investment, agile regulatory frameworks, targeted reskilling programs and strengthened EU digital infrastructure.

Studies and reports

Publications Office
of the European Union