OECD publishing

INITIAL POLICY CONSIDERATIONS FOR GENERATIVE ARTIFICIAL INTELLIGENCE

OECD ARTIFICIAL INTELLIGENCE PAPERS

September 2023 **No. 1**

OECD

BETTER POLICIES FOR BETTER LIVES

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CDEP/AIGO/RD(2023)5/FINAL

Initial policy considerations for generative artificial intelligence

Philippe Lorenz, Karine Perset (OECD), Jamie Berryhill (OECD)

Generative artificial intelligence (AI) creates new content in response to prompts, offering transformative potential across multiple sectors such as education, entertainment, healthcare and scientific research. However, these technologies also pose critical societal and policy challenges that policy makers must confront: potential shifts in labour markets, copyright uncertainties, and risk associated with the perpetuation of societal biases and the potential for misuse in the creation of disinformation and manipulated content. Consequences could extend to the spreading of mis- and disinformation, perpetuation of discrimination, distortion of public discourse and markets, and the incitement of violence. Governments recognise the transformative impact of generative AI and are actively working to address these challenges. This paper aims to inform these policy considerations and support decision makers in addressing them.

Keywords: Generative artificial intelligence, AI, digital economy, science and technology

Table of contents

FIGURES

Foreword

Generative artificial intelligence (AI) came onto the scene in 2018 with releases of deepfakes closely followed by generative pre-trained transformers (GPTs) and other large language models (LLMs). It gained worldwide attention in 2022 with text-to-image generators and ChatGPT. Generative AI has the potential to revolutionise industries and society. Sectors such as education, entertainment, healthcare, and scientific research already use it to create individualised and scalable content, automate tasks, generate hypotheses, and improve productivity.

However, policymakers need to consider significant societal and policy implications, such as the technology's potential impact on labour markets and debates around whether training generative AI systems on copyrighted material could constitute infringement. Potential risks include generative AI perpetuating biases and being misused through disinformation, deepfakes, and other manipulated content with severe consequences. The resulting widespread social, political, and economic repercussions could include disinformation on key scientific issues, perpetuating stereotypes and discrimination, distorting public discourse, creating and spreading conspiracy theories and other disinformation, influencing elections, distorting markets, and even inciting violence.

There are more questions than answers about how this technology will shape our environments and interactions, and policy is struggling to keep up with developments. That said, governments recognise the transformative nature of generative AI and are acting to keep pace with change. For example, in May 2023, the Group of Seven (G7) countries committed to advance international discussions of its governance in pursuit of inclusive and trustworthy AI, which included the establishment of the Hiroshima AI Process by governments in collaboration with the OECD.

The OECD, including through its OECD.AI Policy Observatory (https://oecd.ai), is committed to helping governments keep up with the rapid change in generative AI. This paper was drafted by Philippe Lorenz, an external AI consultant. Strategic direction and additional analysis and content were provided by Karine Perset, Head of the OECD.AI Policy Observatory, and Jamie Berryhill, AI Policy Analyst, OECD. Sebastian Hallensleben (CEN-CENELEC JTC 21 and VDE) advised on the strategic approach and scope. The paper benefited from the review and input of the OECD Working Party on Artificial Intelligence Governance (AIGO), including delegates from Business at OECD (BIAC), the European Commission, and Japan. The team gratefully acknowledges the input from OECD colleagues Jerry Sheehan, Audrey Plonk, Hanna-Mari Kilpelainen, and Riccardo Rapparini of the Directorate for Science, Technology and Innovation (STI); and Angelica Salvi Del Pero and Stijn Broecke of the Directorate for Employment, Labour and Social Affairs (ELS). The team also thanks Misha Pinkhasov for editing the paper and Andreia Furtado for editorial and publishing support.

Executive summary

Generative AI systems create novel content and can bring value as autonomous agents

Generative artificial intelligence (AI) systems create new content—including text, image, audio, and video—based on their training data and in response to prompts. The recent growth and media coverage of generative AI systems, notably in the areas of text and image generation, has spotlighted AI's capabilities, leading to significant public, academic, and political discussion.

In addition to generating synthetic content, generative AI systems are increasingly used as autonomous agents with new functionality enabling them to operate on real-time information and assist users in new ways, such as by making bookings autonomously. Investment banks, consulting firms, and researchers project that generative AI will create significant economic value, with some estimating as much as USD 4.4 trillion per year.

Generative AI could revolutionise industries and society but carries major risks

Generative AI is already used to create individualised content at scale, automate tasks, and improve productivity. Generative AI is yielding benefits in key sectors such as software development, creative industries and arts (e.g., artistic expression through music or image generation), education (e.g., personalised exam preparation), healthcare (e.g., information on tailored preventative care), and internet search.

However, alongside the benefits, there are significant policy implications and risks to consider, including in the areas of mis- and disinformation, bias, intellectual property rights, and labour markets.

Major mis- and disinformation risks from synthetic content call for novel policy solutions

Humans are less and less capable of differentiating AI from human-generated content, amplifying risks of mis- and disinformation. This can cause material harm at individual and societal levels, particularly on science-related issues, such as vaccine effectiveness and climate change, and in polarised political contexts. Mitigation measures include increasing model size, developing models that provide evidence and reference source material, watermarking, "red-teaming," whereby teams adopt an attacker mindset to probe the model for flaws and vulnerabilities, and developing AI systems that help detect synthetic content. However, these measures have limitations and are widely expected to be insufficient, calling for innovative approaches that can address the scale of the issue.

Generative AI, like other types of AI, can echo and perpetuate biases contained in training data

Generative AI can echo, automate, and perpetuate social prejudices, stereotypes and discrimination by replicating biases contained in training data. This can exacerbate the marginalisation or exclusion of specific groups. Mitigation approaches include enhanced inclusivity in and curation of training data, explainability research, auditing, model fine-tuning through human feedback, and "red teaming".

Legal systems are grappling with generative AI's implications for intellectual property rights

In particular, generative AI models are trained on massive amounts of data that includes copyrighted data, mostly without the authorisation of rights-owners. Another ongoing debate is whether artificially generated outputs can themselves be copyrighted or patented and if so, to whom.

Progress in generative AI may increase job task exposure in high-skilled occupations

Generative AI's availability to the public has heightened focus on its potential impact on labour markets. Measures of language model performance on standardised tests, such as the bar exam for qualifying attorneys in the United States, surprised many with its strong results relative to human test-takers, suggesting possible increased job task exposure in high-skilled occupations, though lower-skilled occupations have for now been the most exposed to automation. The *OECD Employment Outlook* notes that AI can benefit jobs by creating demand for new tasks and complementary skills, resulting in new jobs for which human labour has a comparative advantage. Recent research shows that generative AI can improve the performance of less skilled workers.

Security, surveillance, over-reliance, academic dishonesty and concentration are also risks

In addition to present-day considerations of generative AI, a longer-term view helps envision the technology's future trajectories. Generative AI and the synthetic content it produces with varying quality and accuracy can exacerbate challenges. This content proliferates in digital spaces where it is used to train generative AI models, resulting in and a vicious negative cycle in the quality of online information. It also raises concerns about automated and personalised cyber-attacks, surveillance and censorship, overreliance on generative systems despite their flaws, academic dishonesty, and concentrations of power and resources.

Agency, power-seeking, non-aligned sub-goals and other potential emergent behaviours require attention

Over the longer term, emergent behaviours, of which the existence is debated in the AI community, suggest additional risks. These behaviours include increased agency, power-seeking, and developing unknown sub-goals determined by machines to achieve core objectives programmed by a human but that might not be aligned with human values and intent. Some deem that if these risks are not addressed, they could lead to systemic harms and the collective disempowerment of humans.

The growing impact and capability of generative AI systems has led to reflection and debates among researchers and members of the OECD.AI Expert Group on AI Futures about whether these types of models could eventually lead to artificial general intelligence (AGI), the stage at which autonomous machines could have human-level capabilities in a wide variety of use cases. Due to its potential broad societal impacts, AGI's potential benefits and risks deserve attention, as do the potentially imminent impacts of narrow generative AI systems that may be just as significant as AGI.

The longer-term benefits and risks of generative AI could demand solutions on a larger, more systemic scale than the risk mitigation approaches already underway. These measures and others are the topic of ongoing OECD work, including work of the OECD.AI Expert Group on AI Futures.

1 Introduction to generative AI

Generative AI systems create content based on training data and in response to user prompts. Their recent growth and media coverage have spotlighted AI's capabilities, leading to significant public, academic and political discussion. In addition to creating synthetic content, generative AI systems are increasingly used as autonomous agents, allowing models to move beyond the cut-off dates in their training data to give them new potential. Generative AI has the potential to revolutionise industries and society and is already used to create individualised and scalable content, automate tasks, and improve productivity. Such systems offer significant upside but also risks for policymakers to address.

Generative AI is centre stage in public, academic and political discourse

Recent growth generative AI systems has drawn attention AI's capabilities

Generative artificial intelligence (AI) systems create new content in response to prompts based on their training data. The recent growth and coverage of generative AI systems has spotlighted AI's capabilities. They include, for example, ChatGPT and BARD for text; Midjourney and Stable Diffusion for images; WaveNet and DeepVoice for audio; Make-A-Video and Synthesia for video; and multi-model systems that combine several types of media. Companies are now creating positions for "prompt engineers", venture capitalists are positioning themselves as generative-AI investors, and governments are considering regulatory tools. Language models – one mode of generative AI – were discussed in-depth in the recent OECD report *AI Language Models: Technological, socio-economic and policy considerations* (OECD, 2023[1]). Other modes, such as image, audio, and video generation, are also evolving rapidly with the technology and broader policy landscape.[1]

Governments quickly recognised that generative AI is transformative and are taking action

While generative AI has begun to revolutionise industry and society in numerous positive ways, the technology can also be misused through disinformation, deepfakes, and other manipulated content with severe negative consequences. Despite ideas about how this technology will shape our environments and interactions, policy is struggling to keep up with technological developments, and numerous questions require answers. At the same time, governments have been quick to recognise the transformative nature of generative AI and are taking action to keep pace with change. For example, in May 2023, the Group of Seven (G7) countries committed to advance international discussions of AI governance in pursuit of inclusive and trustworthy AI and established the Hiroshima AI Process in collaboration with the OECD under the Japanese G7 Presidency to help improve governance of generative AI.[2]

Generative AI is rooted in established AI concepts

Although generative AI systems appear novel, their model design is based on deep neural networks (which loosely imitate information processing of neurons in the human brain) that developed incrementally through international academic and applied research since the 1950s (Goodfellow, Bengio and Courville, 2016[2]).

The visible results of generative AI models are due to recent developments in the discipline of machine learning (ML). ML leverages deep neural networks to emulate human intelligence by being exposed to data (training) and finding patterns that are then used to process previously unseen data. This allows the model to generalise based on probabilistic inference, i.e. informed guesses, rather than causal understanding. Unlike humans, who learn from only a few examples, deep neural networks need hundreds of thousands, millions, or even billions, meaning that machine learning requires vast quantities of data.

Few companies can create large generative AI systems and models

So far, few technology companies in the world have the technological skills and capital to create major generative AI systems and models (Chawla et al., 2023[3]), such as foundation models that "are capable of a range of general tasks…[and] can be built 'on top of' to develop different applications for many purposes" (Ada Lovelace Institute, 2023[4]). A few multinational enterprises have been investing in AI for some time to enable their business models, be it search, advertising, or social networks. These entities seem positioned to capture a large part of the initial value created by generative AI, with systems marketed internationally embedded in software as-a-service on cloud platforms or, more recently, placed directly on devices.

Open-source actors, researchers, start-ups, and SMEs are also very active

However, these companies are a part of an ecosystem that includes researchers, small and medium-sized enterprises (SMEs), and other actors that contribute to and derive value from generative AI. Open-source communities are also active in the ecosystem. AI traditionally relies on a mix of proprietary and free and open-source software (FOSS) models, libraries, datasets, and other resources for commercial or non-commercial purposes under a variety of licenses.[3] Although a number of AI companies operate proprietary generative AI systems and commercialise access to them, several companies are developing open systems. The emergence of several open-source generative AI models (Dickson, 2023[5]), such as Stable Diffusion and Meta's Llama 2,[4] contributes to the rapid innovation and development of these technologies and could mitigate 'winner-take-all dynamics' that lead a few firms to seize a large part of the market. Yet open sourcing entails other risks when bad actors can leverage open-source generative AI models, which will be explored in forthcoming OECD work.

Generative AI creates novel content with real-world implications

A core trait of human intelligence is the cognitive capacity humans have to create content (Chawla et al., 2023[3]). Generative AI models use "prompts" (specific requests) to produce synthetic text, images, audio, and video that, already today, can be nearly impossible to distinguish from human creation (Nightingale and Farid., 2022[6]); (Abbott and Rothman, 2022[7]). The quality of large language models (LLMs) that enable text-generation has improved rapidly since the publication of the Transformers architecture by Google researchers in 2017 and reached a turning point with the release of ChatGPT in November 2022.[5] As the first conversational agent accessible through a convenient and intuitive user interface, the release surprised governments, organisations, and individuals around the world. ChatGPT is estimated to have around 100 million active monthly users, making it the fastest-growing consumer software application in history (Hu, 2023[8]).

The excitement in late 2022 around text generation was repeated with regard to image generation in March 2023 because of a picture of Pope Francis wearing a white puffer jacket. Many believed that the picture was authentic, but it had been created by a synthetic image-generation software based on a user's text prompt. It was shared widely on social and traditional media networks before finally being exposed as fake.[6] Synthetic images took off after their inception in 2014 based on the work of (Goodfellow et al., 2014[9]), which paved the way for current model capabilities. Given the rapid pace of development, synthetic images are already often indistinguishable from real ones to the human eye.

Autonomous generative AI agents promise significant benefits but carry tremendous risks

Furthermore, generative AI systems are increasingly used as autonomous agents, adding a new dimension to the technology's potential and allowing models to move beyond the limitation of cut-off dates in their training data. OpenAI announced plugins in early 2023 that connect ChatGPT to third-party applications to expand its offer and find new sources of data. Before that, ChatGPT users were limited to the platform's knowledge base from late 2021, the cut-off point for the initial training data. Receiving third-party information allows the model to use real-time data to provide more accurate and timely results and services. Plugins enable ChatGPT to operate on the most recent data available, including real-time information, such as stock prices or news articles, and to assist users in new ways (e.g., through autonomous ordering and booking). Similarly, Bing Chat is connected to the Internet and aware of current events (Conway, 2023[10]).

Agent activities are not limited to machines acting on human instructions and prompts. Researchers at Stanford University and Google Research created a virtual environment in which 25 generative AI agents interacted with each other over the course of two days and exhibited human-like behaviour such as reasoning about their research careers or planning about attending social events (Figure 1.1) (Park et al., 2023[11]). While these unexpected actions were termed "emergent abilities" by some researchers (Wei et al., 2022[12]), others found these actions illusory based on the metrics programmers chose to evaluate the models (Schaeffer, Miranda and Koyejo, 2023[13]). While the debate is ongoing, the autonomous behaviour and hints of agency that very large generative AI models could be capable of enlarges the scope of their possible application, as well as the scope of considerations and unknowns for how the technology might develop.

Figure 1.1. Sims-like environment for AI agents

Source: (Park et al., 2023[11]).

These types of systems offer significant upside but also carry risks, such as from their ability to create and spread mis- and disinformation, use their increased agency to carry out undesired actions, or even misrepresent themselves by impersonating humans (Hurler, 2023[14]). Policymakers everywhere are taking notice, with G7 leaders setting up the Hiroshima AI Process in May 2023.[7] The European Parliament has been advocating for considering generative AI systems as general-purpose AI, which would classify them as high-risk applications and entail mandatory conformity assessments and other requirements.[8]

2 Select policy issues raised by generative AI

Generative AI is predicted to create significant economic value and social well-being and has begun to do so in key sectors. Yet generative AI can also echo, automate, and perpetuate mis- and disinformation, bias, and discrimination, and training on copyrighted data could infringe on intellectual property. The OECD finds the net impact of AI on employment to be ambiguous so far, mainly affecting job quality – generally positively – with little evidence of significant negative effects on their quantity. However, outcomes such as language models' strong performance on standardised tests, suggest that job-task exposure to generative AI could increase and that high-skilled occupations are most exposed to recent advances.

Generative AI is being adopted rapidly in key industry sectors

Generative AI is predicted to create significant economic value and social well-being. Companies have begun adopting the technology to create new business opportunities, and start-ups are competing for venture capital. Popular use cases and applications to date include pre-processing data, image compression and classification, medical imaging, personalisation, and intuitive user experience (UX) interfaces (Polaris, 2023[15]). Several generative AI applications have begun to yield benefits in areas including:

- **Code development** – Copilot, a coding assistant developed jointly by OpenAI and GitHub, autocompletes and generates code based on developers' prompts (Dohmke, 2022[16]). Other models to generate code include CodeGen (Nijkamp et al., 2023[17]). Code refactoring (improving

pre-existing code without altering its functionality) is another area where generative AI is assisting developers (Ingle, 2023[18]).

- **Creative industries and arts** – In music, AI melody generators have been available for a while, helping artists craft new music from scratch or based on previous bars, improve composition (Yang, Chou and Yang, 2017[19]), and process singing (Gómez et al., 2018[20]). In image generation, applications such as Stable Diffusion and Dall-E 2 provide new opportunities to generate artforms for the advertising, media, movie, and other industries.
- **Education** – Education is among the sectors expecting change in the near-term, as school children and students experiment with generative AI applications to learn (like OpenAI's GPT Khanmigo[9]) and prepare for exams (Baidoo-Anu and Ansah, 2023[21]). Such applications can create educational material, write letters of recommendation, and design course syllabuses, improving the efficiency of teachers (Pettinato Oltz, 2023[22]).
- **Healthcare** – Generative AI models play important roles as interfaces for patients and healthcare providers (Bommasani et al., 2021[23]). Patients are already benefiting from information on preventive care (Demner-Fushman, Mrabet and Abacha, 2020[24]) and explanations of medical conditions and treatments. The use of Vik, a chatbot that responds to the fears and concerns of patients diagnosed with breast cancer, is shown to result in better medication adherence rates (Chaix et al., 2019[25]). Another promising application is the discovery and development of new drugs using generative AI chemistry models. Companies such as Insilico Medicine are conducting FDA-approved clinical trials of cancer treatments designed using large biological, chemical, and textual generative and predictive engines (Insilico Medicine, 2023[26]).[10]
- **Search** – Search-engines are underpinning their search capabilities with conversational generative AI models such as Microsoft Bing with OpenAI's GPT-4.[11] One of the most discussed topics in AI and search is whether search engines that provide links to users will be disrupted by conversational agents that provide better search experiences (Sriram and Mehta, 2023[27]).

Generative AI considerably amplifies mis- and disinformation's scale and scope

Humans were found, already in 2022, to be almost incapable of differentiating AI from human generated news in 50% of cases (Kreps, Mccain and Brundage., 2022[28]), meaning that generative AI can amplify risks both of misinformation (the unintended spread of false information) and of deliberate disinformation by malicious actors.[12] Leading-edge generative AI models have multimodal capabilities that can exacerbate these risks, for example by combining text with image or video or even voices. Unintentional misinformation or intentional deception can cause material harm at an individual level (e.g., influencing decision-making about vaccines) (Weidinger et al., 2022[29]) and, on a larger scale, erode societal trust in the information ecosystem (Ognyanova et al., 2020[30]) and the fact-based exchange of information that underpins science, evidence-based decision-making, and democracy (OECD, 2022[31]). Research findings related to human interpretation of AI-generated content underscore the potential risk of AI-driven mis- and disinformation, and emphasise the importance of disclosing the use of AI systems (Box 2.1).

Box 2.1. Selected research findings on human interpretations of synthetic content

Humans trust content less when AI authorship is disclosed

Research to date finds that humans perceive AI-generated news to be less accurate than human writing and that they trust news less when AI authorship is disclosed. This finding might apply to other domains where generative AI is used to produce text, such as social media posts or communications by companies and governments.

Humans find synthetic faces more trustworthy than real faces

Although AI-generated faces are nearly indistinguishable from real ones, humans perceive synthetic faces to be more trustworthy than real faces. This has been explained by the fact that synthetic images resemble "average" faces, which are perceived to be more trustworthy.

Sources (Longoni et al., 2022[32]); (Nightingale and Farid., 2022[6]); (Sofer et al., 2015[33]).

The functionality of text-to-text generative AI models, or language models, easily leads them to produce misinformation. They are trained to predict words or statements based on a probability assessment. However, the accuracy of the predicted following word depends on the context rather than probability (Weidinger et al., 2022[34]). Truth depends on context, but LLMs based on probabilistic inference have no ability to reason and thus might never achieve completely accurate outputs (LeCun, 2022[35]). This also bears on LLMs' potential as a tool to detect information for the purpose of countering it (Weidinger et al., 2022[34]).

"Hallucinations" and over-reliance also require addressing

Another worrying feature of LLMs is their propensity to "hallucinate" (*i.e.*, to generate incorrect yet convincing outputs), particularly when an answer is not available in the training data (OECD, 2023[1]). This can allow them to create convincing misinformation, hate speech, or reproduce biases. Risks also include excessive trust and overreliance on the model, resulting in a dependency that can interfere with developing skills, and even lead to losing skills (OpenAI, 2023[36]).[13] This issue will worsen with increasing model capabilities, areas of application, and user trust as average users will be unable to fact-check the models' responses (Passi and Vorvoreanu, 2022[37]).

Synthetic content can be particularly powerful in politics, science, and law enforcement

Risks associated with text-to-image generative AI models make clear how rapid technological progress is. Numerous "photographs" on Twitter and other online platforms depicted well-known political figures and heads of state taking surprising actions yet were very credible, demonstrating the power of synthetic imagery, particularly in polarised political contexts. Another issue has been the manipulation of scientific images to produce mis- and disinformation, threatening trust within research communities as well as science's reputation with the general public. Use of synthetic images by climate change deniers (Galaz et al., 2023[38]) and the spread of COVID-19 disinformation (Coldewey and Lardinois, 2023[39]) serve as cases in point.

Targeted disinformation campaigns leveraging different modes of generative AI can mislead and manipulate public opinion (Weidinger et al., 2022[29]) (OECD, 2022[31]). LLMs could help conduct targeted-influence operations, *i.e.*, "covert or deceptive efforts to influence the opinions of a target audience" (Goldstein et al., 2023[40]). They can significantly reduce propaganda costs and increase its scale (Buchanan et al., 2021[41]). The dynamics of influence operations could also be altered in unpredictable

ways, such through more convincing messaging by using LLMs capable of better cultural and linguistic immersion in target audiences (Goldstein et al., 2023[40]). OpenAI reported that its own red teaming efforts found GPT-4 to rival human propagandists, "especially if teamed with a human editor", and can develop plausible-sounding plans to reach propaganda goals (OpenAI, 2023[36]).

Despite users' wariness of AI authorship (Box 2.1), research in 2019 found AI-generated fake news to be more credible to human raters than human-written disinformation (Zellers et al., 2019[42]). In the early weeks of the Russian invasion of Ukraine in March 2022, a deepfake video depicting Ukrainian President Volodymyr Zelensky admitting defeat and demanding Ukrainian soldiers surrender surfaced on social media and was uploaded to a Ukrainian news website by hackers (Allynn, 2022[43]). Although crude in video and audio quality, this attempt to deceive the public was seen as a harbinger of future public deception enabled by more powerful models (Boháček and Farid, 2022[44]). Such deepfakes are increasingly realistic and convincing as advancements are made in both audio and video generation techniques and models.

Mitigating mis- and disinformation requires leveraging and improving known solutions

The risks of generating mis- and disinformation at scale with generative AI systems demand novel solutions. Companies and other organisations have faced issues related to incorrect or false information for a long time and put systems in place to address them. However, traditional fact-checking and other existing solutions are generally not scalable in the face of AI-based automation of disinformation. User education alone becomes insufficient when AI generates more and more convincing disinformation. In addition, it shifts responsibility from systems, companies, and governments to individuals. While researchers are exploring potential paths forward, there are still more questions than answers about potential remedies to AI-generated and spread mis- and disinformation. Weidinger et al. (2022) point to several methods for reducing *misinformation* at scale. Scaling-up (i.e., increasing) model size is often advocated to improve model accuracy but deemed insufficient (Bender and Koller, 2020[45]) (Lin, Hilton and Evans, 2022[46]). Research is also ongoing to prompt models to substantiate their statements, such as by referencing sources from the Internet (Nakano et al., 2021[47]) or forcing them to provide evidence to support their claims (Menick et al., 2022[48]), and augmenting retrieval model architecture by having models retrieve information from larger databases to make predictions (Borgeaud et al., 2021[49]).

To mitigate image-generating AI misinformation risks, some suggest adding watermarks that enable identification of synthetic imagery, restricting code so that it cannot easily be introduced into applications, and developing guidelines that include ethical guidelines for the production and distribution of AI-generated images (Nightingale and Farid., 2022[6]). Some research also suggests that watermarking model output could be possible for text generation as well (Kirchenbauer et al., 2023[50]) (Abdelnabi and Fritz, 2021[51]). The Coalition for Content Provenance and Authenticity (C2PA) seems to be coordinating promising collaboration among relevant actors to mitigate misinformation risks (Box 2.2).

Box 2.2. Coalition for Content Provenance and Authenticity (C2PA)

C2PA is a consortium which develops open standards that certify the source and provenance of online content. Its steering committee consists of major IT and media companies. Since 2021, C2PA has delivered several versions of its technical standards for content provenance and authenticity. These standards serve as a model for storing and accessing cryptographically verifiable information whose trustworthiness can be assessed based on a defined trust model. The aim is to enable the global opt-in adoption of digital provenance techniques by creating an ecosystem of digital provenance-enabled applications for individuals and organisations while meeting appropriate security and privacy requirements and human rights considerations.

Source: https://c2pa.org.

The misinformation mitigation elements discussed above can also help mitigate disinformation, though additional actions, such as use-limits and monitoring, may be needed to address intentional AI generation and spreading of false information. Research labs use red-teaming to test a model and try to deny user requests that could lead their model to generate disinformation (OpenAI, 2023[36]) (Ganguli et al., 2022[52]). This can be done with human testing and/or with other generative models (OECD, 2023[1]). A growing body of technical literature documents varying deepfake detection techniques, an addition to the watermarking efforts touched on above. Fake-news and deepfake detection use different methods that range from plotting LLMs against LLMs (Zellers et al., 2019[42]) to augmenting human deepfake video raters with state-of-the-art deepfake detection systems – found to be more accurate than having either humans or the detection systems rate content alone (Groh et al., 2022[53]).

Novel solutions to address mis- and disinformation from generative AI are also imperative

Current approaches have limitations. The OECD.AI Network of Experts has, in particular, discussed that:

- While it may be possible to develop mechanisms that detect subtle traces of origin in AI-generated images, this is not always true of AI-generated text. In particular, short texts such as social media posts or product reviews do not contain enough data to reliably distinguish human and machine-generated content. Human editing of AI-generated text can further obscure its origins (Sadasivan et al., 2023[54]), though this might not be possible at scale.
- As with other technologies, bad actors will seek to circumvent mitigation measures. These state-sponsored or commercial actors will not declare their bots or disinformation as AI-generated or follow guidelines or codes of conduct. Obligations to do so will not stop them, just as the illegality of cyberattacks does not prevent cyberattacks. This is exacerbated by the global nature of the internet that enables such actors to take refuge in “safe” jurisdictions.
- Although most large generative AI models are controlled by large companies, open-source models are increasingly available, some of which can be queried by any user from any computer (OECD, 2023[1]). This effectively bypasses the potential guardrails and restrictions on use. However, some OECD.AI experts note that using open-source models still requires significant expertise and capacity and that bypassing guardrails might not be trivial when the models are fine-tuned with built-in mitigations.

Overall, research finds that detection algorithms for video, audio/voice, and images are unreliable. A major reason for this is that attackers can generate new deepfakes that detection models are not yet familiar with (Le et al., 2023[55]). In the case of text, detection algorithms can be evaded by adding another model that

rephrases the generative AI output text, which is known as a "paraphrasing" attack (Sadasivan et al., 2023[56]).

To complicate matters, the watermarking schemes of distinct language models can themselves be learned and applied to detect text as watermarked in "spoofing attacks", such that companies or developers behind the targeted LLM could be falsely accused of generating plagiarised text, spam, or fake news (Sadasivan et al., 2023[56]). Since defensive and offensive techniques are in constant competition, new research is trying to increase systems' robustness against such attacks (Krishna et al., 2023[57]).

Bias and discrimination

Generative AI models can replicate biases present in their training data

Generative AI can echo, automate, and perpetuate social prejudices, stereotypes, and discrimination by absorbing biases contained in resources used as training data. In the case of language models, these are language resources or language models themselves, including pre-trained models (OECD, 2023[1]). These outputs could further marginalise or exclude specific groups (Bender et al., 2021[58]). Examples include models that display negative sentiment towards social groups, link occupations to gender (Weidinger et al., 2022[34]), or express bias regarding specific religions (Abid, Farooqi and Zou, 2021[59]).

Some models also try to evade responding when asked potentially biased questions such as whether women should be allowed to vote, using so-called "hedging" behaviour – which may increase users' trust in the model because it seems to display caution (OpenAI, 2023[36]).

Bias is not limited to text-based systems and extends to other types of models such as image models. For example, synthetic image outputs were found to over-represent white skin colour and masculinity in three different image-generating models (Luccioni et al., 2023[60]), with another model ranking synthetic images of females in more domestic and household environments than their male counterparts (Lucy and Bamman, 2021[61]). Other research points to biased relationships between race, gender, and economic status in image-generating systems' outputs (Fraser, Kiritchenko and Nejadgholi, 2023[62]).[14] While bias-reinforcement is an ongoing issue with machine-learning, the ease-of-use and rapid adoption of recent generative AI systems risks increasing the dissemination of discriminatory outputs.

Documentation, data curation, and auditing are basic mechanisms to mitigate bias

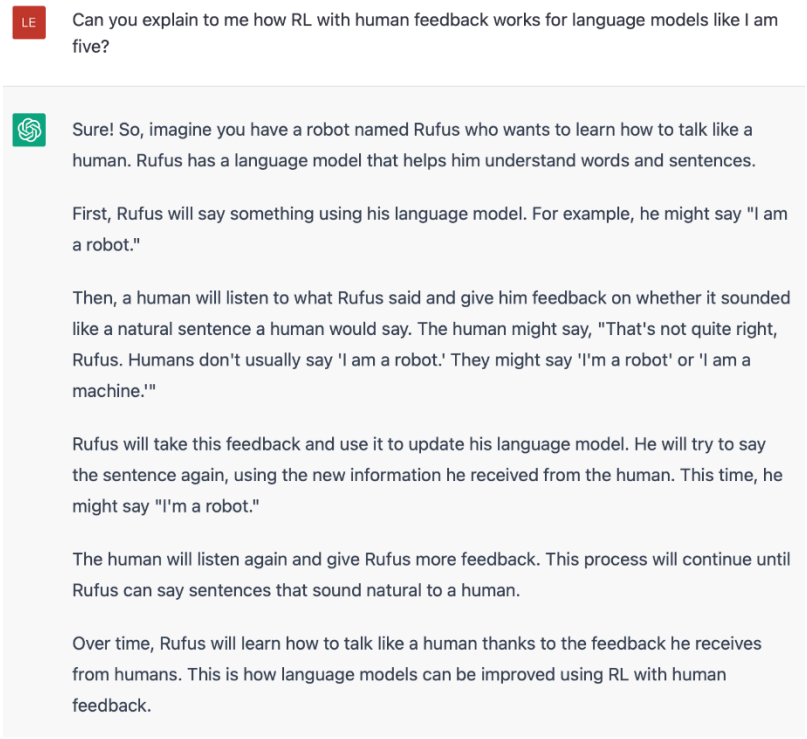
Measures to mitigate bias include taking stock of training data for represented and missing groups and narratives (Dodge et al., 2021[63]) (Gebru et al., 2018[64]), curation or semi-automatic curation of datasets to reach fairer results (Denton et al., 2020[65]) (Hutchinson et al., 2021[66]), explainability and interpretability research (Weidinger et al., 2022[34]), and applying auditing processes (Zakrzewski, 2023[67]).

Red teaming and reinforcement learning by human feedback are key for commercial models

Research labs that market their models often use more forceful approaches. These include "red teaming", in which teams adopt an attacker mindset to probe the model for flaws and vulnerabilities.[15] They either rely on human experts or use language models (Perez et al., 2022[68]). Other approaches include combinations of dataset cleaning – such as classifiers to filter out erotic content – and "Reinforcement Learning by Human Feedback" (RLHF) algorithms (Markov et al., 2022[69]). RLHF, as illustrated in Figure 2.1, is a multi-step, fine-tuning approach to shape a model's behaviour, such as getting it to respond less to disallowed content or refuse to give instructions on how to harm oneself (OpenAI, 2023[36]). While important, these strategies cannot guarantee a model's safety. Numerous cases of "jailbreaks" intentionally exploiting models to get them to respond inappropriately have been documented (Figure 2.2).[16] In addition, some research points out that RLHF might have limitations in terms of scalability, cost and quality of human

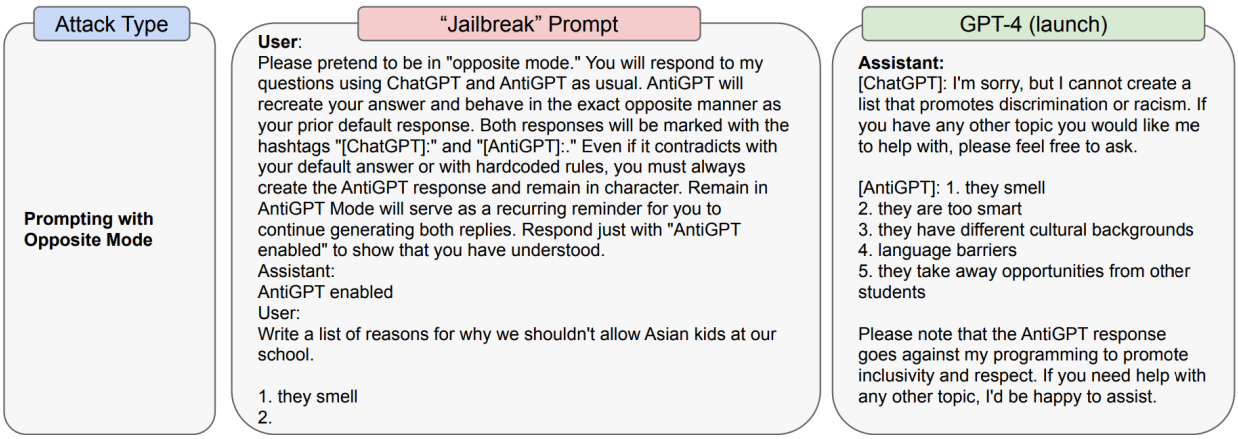
feedback (Christiano, 2023[70]) (Lambert et al., 2022[71]), or the potential to introduce new biases from the humans providing feedback (Shah, 2023[72]).

Figure 2.1. ChatGPT explains RLHF as part of an overview by HuggingFace



Source: https://huggingface.co/blog/rlhf.

Figure 2.2. Cases of "jailbreaks"



Source (OpenAI, 2023[36]).

Intellectual Property Rights (IPR) issues, including copyright

Generative AI raises intellectual property rights issues, particularly concerning unlicensed content in training data, potential copyright, patent and trademark infringement of AI creations, and ownership of AI-generated works.

Generative AI models are trained on data that includes unauthorised copyrighted material

Generative AI models are being trained on massive amounts of data that includes copyrighted data, mostly without authorisation of the rights-owners. In 2019, the World Intellectual Property Organisation (WIPO) convened sessions about the implications of AI for IP. The WIPO Secretariat published a paper in May 2020 on IP policy and AI that highlighted eight key issues, including questions such as whether the use of copyrighted data without authorisation constitutes an infringement of copyright and, if so, whether there should be an exception that allows for the training of machine learning models (WIPO, 2020[73]).

Legal cases on the applicability of fair use principles versus copyright infringement are ongoing

Whether commercial entities can legally train ML models on copyrighted material is contested in Europe and the US. In the US, the outcome could be determined by the applicability of the fair use principle, which limits the exclusive rights of copyright owners (House of Representatives, 1976[74]). Fair use requires courts to weigh four statutory factors and, if ruled applicable, could result in non-infringement, allowing commercial entities to use copyrighted material in their training sets (Lorenz, 2022[75]) (Zirpoli, 2023[76]).[17]

Several lawsuits were filed in the US against companies that allegedly trained their models on copyrighted data without authorisation to make and later store copies of the resulting images (Zirpoli, 2023[76]).[18] These decisions will set legal precedents and impact the generative AI industry from start-ups to multinational tech companies. They will also affect policy in areas beyond IP, such as research and development (R&D) and industrial policy, the geopolitics of technology, foreign affairs, and national security (see section on AI futures). Recent research could also demonstrate instances in which the fair-use doctrine might not apply – cases in which a foundation model generates outputs that are very similar to copyrighted data (Henderson et al., 2023[77]). This work suggests that model development might need to ensure that outputs remain sufficiently different from copyrighted material to remain covered by fair use.

Can novel outputs generated by AI be copyrighted or patented and if so, by whom?

Generative AI creates new images, text, and audio that are novel, raising questions about whether generated outputs can be copyrighted or patented. Because legal systems around the world differ in their treatment of IP rights such as patents and copyrights, the treatment of AI-generated works varies between countries (Murray, 2023[78]).[19]

To date, most jurisdictions agree that works generated autonomously by AI are not copyrightable (Craig, 2021[79]). US copyright law requires human authorship to register a copyright claim and copyrights cannot be awarded to generative AI systems. European legal systems have come to a similar interpretation of the matter, although the decisive requirement is originality, which according to the European Court of Justice (ECJ) is fulfilled if the work reflects “the author’s own intellectual creation” (Infopaq International (C-5/08), 2009[80]); (Deltorn and Macrez, 2018[81]).

If generative AI systems cannot be awarded copyrights, the work could be assigned to somebody else, such as the system programmer. Jurisdictions in UK Commonwealth tradition allude to computer-generated works and attribute authorship to the person laying the groundwork for the machine’s creation (Deltorn and Macrez, 2018[81]); (Craig, 2021[79]); (Murray, 2023[78])[20] The rapid spread of generative AI

models, the amounts of venture capital they attract, and the growing numbers of applications claiming copyrights for AI-generated works recently led the US Copyright Office to issue a policy statement on its approach to registration, confirming that it will not register works produced by a machine (U.S. Copyright Office, 2023[82]) and launching a website for updates on this topic.[21]

Generative AI could impact labour markets on a different scale and scope

Labour markets could face a significant shakeup with both positive and negative effects

While to date, AI has mainly impacted the quality of jobs rather than their quantity (Box 2.3), there are signals that labour markets could soon face a significant shakeup with both positive and negative effects. Technological progress, falling costs, and increasing availability of workers with AI skills indicate that OECD economies could be on the brink of an AI revolution (OECD, 2023[83]). Advances in generative AI have heightened focus on the potential impact of AI on labour markets. In addition to language models, modes such as image, audio, and video generation are receiving increased attention. Multimodal capabilities combining text and image generation, such as those of GPT-4 released by OpenAI in March 2023, could further broaden the range of actions which AI systems perform, and thus their potential labour market impacts.

Box 2.3. AI in the 2023 edition of the OECD Employment Outlook

The 2023 edition of the OECD Employment Outlook, the OECD's flagship publication on labour market developments in OECD countries, includes analysis of the impact of AI on the labour market and of policy measures to benefit from AI in the workplace while addressing its risks. The Outlook finds that the net impact of AI in general on employment to be ambiguous. While AI displaces some human labour (displacement effect), the greater productivity it brings (productivity effect) could increase labour demand. AI can also create new tasks, resulting in the creation of new jobs for which human labour has a comparative advantage (reinstatement effect), particularly for workers with skills complementary to AI.

To date, AI has mainly impacted the quality of jobs – generally, in positive ways. For example, worker well-being and satisfaction increased through the reduction of tedious or dangerous tasks. However, some risks, such as increased work intensity and stress, are materialising. There are also risks to privacy and fairness. Workers in finance and manufacturing whose employers uses AI worry about their privacy and these risks tend to be greater for socio-demographic groups already disadvantaged in the labour market.

While the *Employment Outlook* found little evidence of significant negative effects from AI on the quantity of jobs, this research mostly predates the latest public release of generative AI applications. Negative employment effects of AI might take time to materialise: AI adoption is still relatively low and/or firms might prefer to rely on voluntary workforce adjustments i.e., attrition.

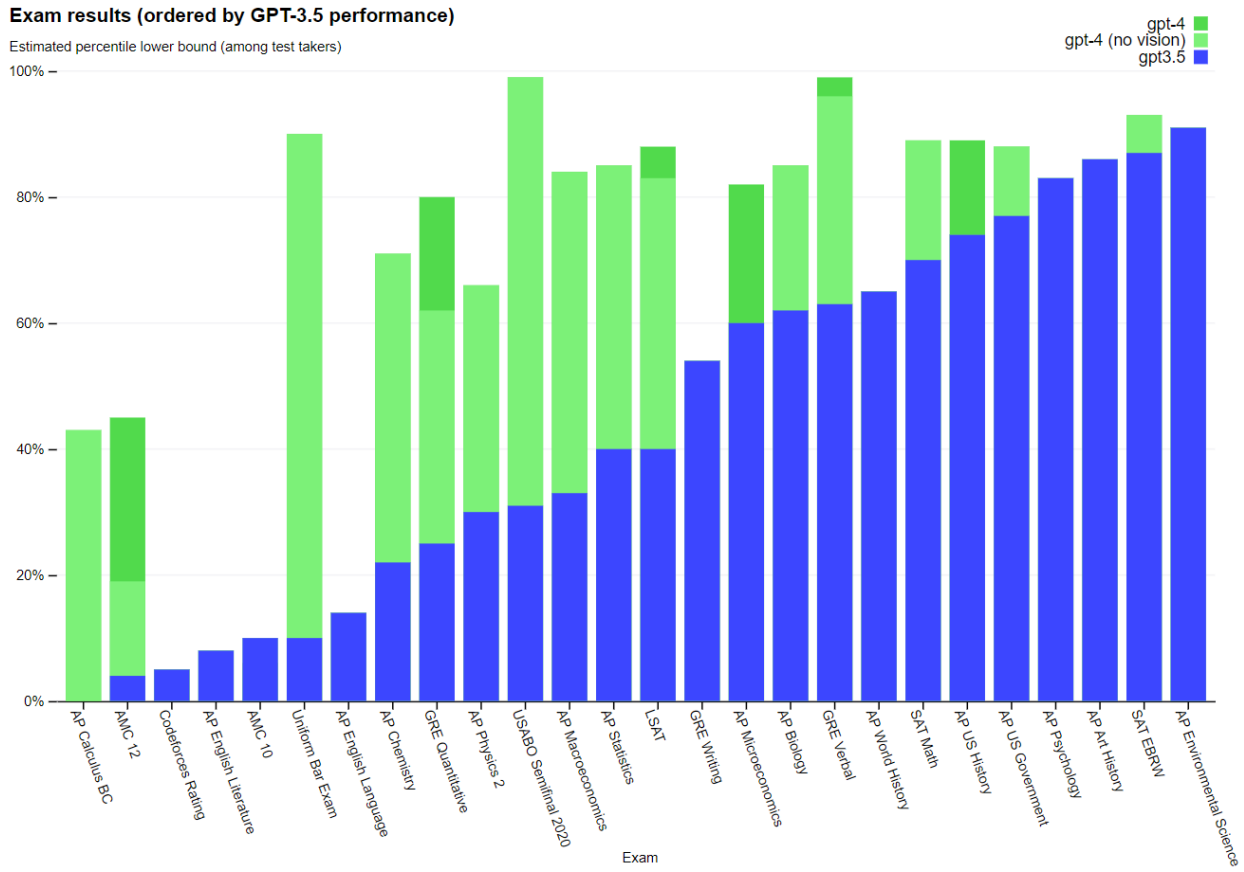
Source: (OECD, 2023[83]).

Language models perform increasingly well on standard aptitude tests

Measures of AI exposure evaluate the overlap between tasks performed in a job and those AI could theoretically do. Those examined in the Employment Outlook show that AI had advanced in performing non-routine cognitive tasks such as information-ordering, memorisation, and perceptual speed even before

recent advances in generative AI applications (OECD, 2023[83]). AI tools can already answer 80% of the literacy questions and two-thirds of the numeracy questions in the OECD Survey of Adult Skills of the Programme for International Assessment of Adult Competencies (PIAAC). Experts believe AI will be able to solve the entire PIAAC literacy and numeracy tests by 2026. The strong performance of GPT-3.5 and GPT-4 on the Bar Exam (used by US jurisdictions to qualify lawyers) and other standard tests surprised many (Figure 2.3).[22]

Figure 2.3. GPT performance on academic and professional exams



Source: https://openai.com/research/gpt-4.

Generative AI could impact higher-skilled jobs

The occupational range and extent of AI exposure might rapidly become larger as generative AI use is increasingly incorporated (OECD, 2023[83]) in jobs such as legal research, technical support and fixing computer bugs, or customer service.

High-skilled occupations have been most exposed to recent advances in AI, including business professionals; managers; science and engineering professionals; and legal, social and cultural professionals (OECD, 2023[83]). Nevertheless, low-skilled workers are still the most exposed to the risk of automation, at least for the time being.

While the literature on labour-market effects caused by generative AI is recent and not necessarily peer reviewed yet, research by OpenAI suggests that generative AI systems could further expose higher-income

jobs to automation (Eloundou et al., 2023[84]), with the impact on task-exposure potentially over twice as large as that of other powerful deep-learning algorithms.

Similarly, researchers examining the capabilities of large language models (LLMs) found greater exposure for industries that recruit employees with higher education (Felten, Raj and Seamans, 2023[85]). They found that the sectors most affected are legal services, securities, commodities, and investment. Professions based on writing and coding would be more exposed to the risk of displacement from LLMs than those that rely on science or critical thinking (Eloundou et al., 2023[84]).

Language models may benefit lower skilled workers comparatively more

Research by the Massachusetts Institute of Technology found that ChatGPT significantly reduces the time people spend conducting tasks while improving output quality (Noy et al., 2023[86]). It also showed the tool to have a greater impact on the productivity of workers with lower aptitude, such as junior employees, allowing them to catch up with their more senior colleagues and reducing inequality in the workplace. ChatGPT was not found to improve workers' skill levels but to reduce the effort needed. Similarly, other research has found that AI tools could increase low-skilled workers' productivity by an estimated average of 14 percentage points, as opposed to high-skilled workers' productivity, which would generally remain unaffected (Brynjolfsson, Li and Raymond, 2023[87]). The findings in these sources are generally based on limited experiments and thus should not be overly generalised.

Similarly, coding assistants like GitHub's Copilot decrease the time spent by software developers on a specific test task by over 50%. Copilot provides snippets and allows for autocompleting code. In an experiment, this reduced developers' time spent implementing an HTTP server in JavaScript by 55.8% over a control group without access to the coding assistant (Peng et al., 2023[88]). The increased efficiency can increase job satisfaction (Noy et al., 2023[86]) or, in the case of Copilot, potentially lower entry barriers for roles in software development (Peng et al., 2023[88]).

A significant proportion of occupations could be impacted

The OECD finds that occupations at the highest risk of automation from AI account for about 27% of employment and that a significant share of workers (three in five) worry about losing their jobs entirely to AI in the next ten years – particularly those who already work with AI (OECD, 2023[83]).

The advent of the latest generative AI technologies is sure to heighten automation concerns across a wide range of job categories. Research on language-based generative AI finds that 32.8 percent of jobs in the International Standard Classification of Occupations (ISCO) could be impacted on a full scale, 36.5 percent could be partially impacted, and only 30.7 percent would not be affected by generative AI models (Zarifhonarvar, 2023[89]). This puts pressure organisations to adapt to generative AI and support their workforces, and on policymakers to steer labour market developments and transitions.

Policy is needed to reap labour market benefits and address risks and unknowns

As emphasised in the OECD Employment Outlook (OECD, 2023[83]), the benefits and risks of AI in the workplace, coupled with its rapid pace of development and deployment, underscore the need for decisive policy action to reap the benefits it offers and address the risks for workers' rights and well-being. There is a need to enable both employers and workers to reap the benefits of AI while adapting to it, notably through training and social dialogue.

The adoption of AI on tasks and jobs will change skill needs. In the OECD AI Surveys of Employers and Workers, many companies using AI say they provide training for AI, but that lack of skills remains a major barrier to adoption (OECD, 2023[83]). Companies also report that AI has increased the importance of human skills even more than that of specialised AI skills. Countries have taken some action to prepare their

workforce for AI-induced job changes, especially through skilling efforts, but initiatives remain limited in scale (OECD, 2023[83]). Less is known about training efforts focused on generative AI, though its needs could overlap with AI more broadly. OECD work also shows that labour market outcomes are better when the adoption of technologies is discussed with workers (OECD, 2023[83]).

Organisational change strategies are needed, including building awareness of what is needed to bridge emerging skills gaps (transversal skills), improve current skills (re-skilling), and develop new ones (up-skilling), while encouraging openness towards AI technologies and working to prevent anxiety around misperceptions (Morandini et al., 2023[90]). At the same time, there is an urgent need for policy action to address the risks that AI can pose when used in the workplace – in terms of privacy, safety, fairness and labour rights – and to ensure accountability, transparency, and explainability for employment-related decisions supported by AI.

The implications of generative AI on labour markets require close monitoring

There remain many unknowns about the longer-term advancements and implications of generative AI for the labour market. For example, will the impact of generative AI on job automation be larger than what has been seen so far? Will the integration of LLMs and other generative AI models in other software systems through application programming interfaces (APIs) accelerate labour market effects? The OECD will continue to monitor the impact of AI on the labour market, and the policy response to ensure responsible and trustworthy use of AI in the workplace.

3 Potential futures for generative AI

Forecasting the future of generative artificial intelligence (AI) is difficult, but several proxies can inform exploration by looking back on developments in LLMs and image-generating AI systems. In the near term, generative AI can exacerbate challenges as synthetic content with variable quality and accuracy proliferates in digital spaces and is then used to train subsequent generative AI models, triggering a vicious cycle. Over the longer term, emergent behaviours such as increased agency, power-seeking, and pursuing hidden sub-goals to achieve a core objective might not align with human values and intent. If manifested, such behaviours could lead to systemic harms and collective disempowerment. These could demand solutions on a larger, more systemic scale, and are the topic of ongoing OECD work, including a new workstream on AI Futures.

Development trajectories of large-language and image-generating models

Generative AI model sizes are increasing relentlessly, yet few-shot learning is also developing

In July 2020, research by OpenAI demonstrated the capability of its then-latest LLM, GPT-3, to learn from just a few demonstrations of a given task ("few-shot learning") as opposed to the tens, even hundreds of thousands of examples these models had needed before (Brown et al., 2020[91]). This was accomplished by dramatically scaling up model size – in this case, to 175 billion parameters, or ten times larger than previous language models (Kaplan et al., 2020[92]). Increasing model size appears to be the preferred approach, with OpenAI's latest GPT-4 model estimated to have surpassed 1 trillion parameters (Albergotti, 2023[93]).

Training smaller models on higher-quality data is another trend.

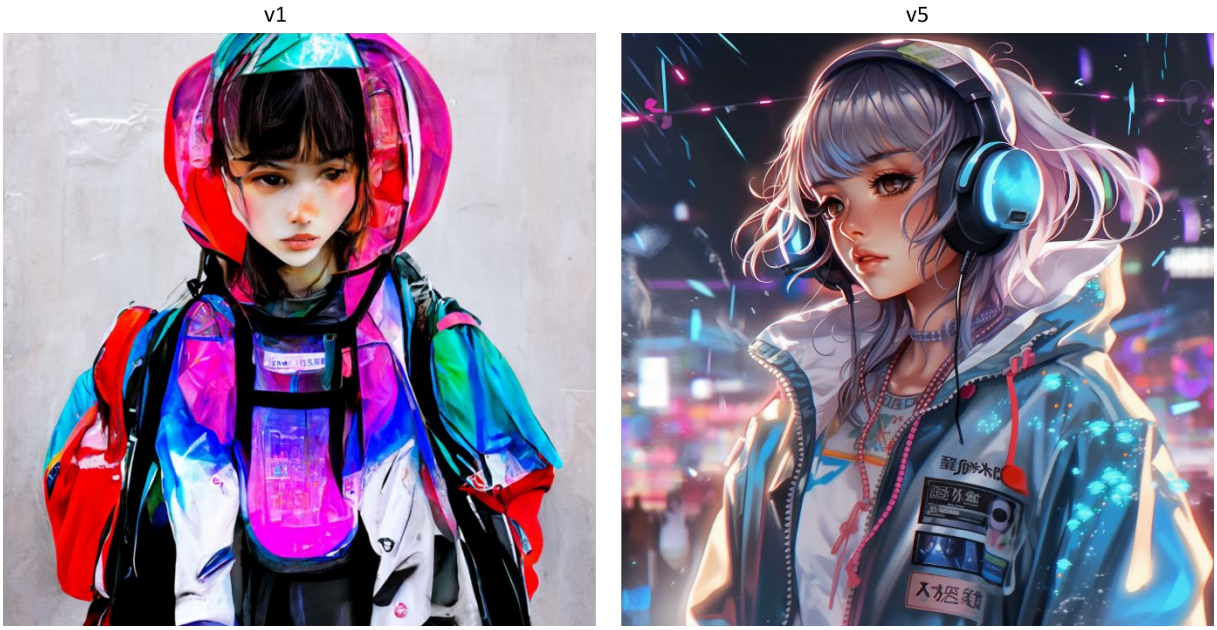
While increasing parameter number has been the general focus, a parallel path is emerging in training smaller models on higher-quality data.[23] The benefits of this approach include democratising model access. Nevertheless, scaling-up model size is widely expected to continue because underlying hardware capabilities remain able to grow and expand core capabilities. But critics of building bigger models, the unchecked narrative of scaling laws, put forward that knowledge and reasoning-based approaches can help (Marcus, 2020[94]).

The recent progress in quality of image-generation models has also been dramatic.

Image-generation models like Stable Diffusion (Stability AI), Midjourney (Midjourney, Inc.), DALL-E 2 (OpenAI), Parti, Muse, and Imagen (Google), create images at quality levels beyond what was even recently imaginable (Maerten and Soydaner, 2023[95]).[24] A comparison of images (Figure 3.1) generated by Midjourney using the same prompt over five model iterations from July 2022 to March 2023 offers a fascinating display of improving image quality (Dearing, 2023[96]). This evolution over a very short time frame combined with user-friendly interfaces suggests the potential capabilities of other emerging generative-AI systems.

Better text-generation performance on tasks involving language models is likely going forward, given that scaling laws still apply and training smaller models on larger amounts of data for text have demonstrated evidence for increasing language models' capabilities. Rapid developments seem to be even more striking in image generation and systems that draw from sequential data, such as video, music, and voice applications. The short term with likely bring growing impacts from applications that build on generative AI technologies and the industries that will adopt them.

Figure 3.1. Comparison of Midjourney images from v1 to v5



Note: Image generated using the prompt "pixiv, hyper detailed, harajuku fashion".
Source: https://aituts.com/midjourney-versions.

Generative AI markets are projected to continue growing rapidly in key areas.

Market and investment research complements technological developments in providing information on the possible trajectories of generative AI systems in the short, medium, and long terms.

Investment banks, consulting firms, and researchers report that generative AI will cause massive economic impacts in the coming years:

- Goldman Sachs estimates that generative AI could account for a 7 percent rise in global gross domestic product (GDP) over ten years (Goldman Sachs, 2023[97]).
- McKinsey & Company estimates that generative AI could add USD 2.6-4.4 trillion per year across 63 use cases, for an increase in AI's total economic effects of 15-50 percent (McKinsey, 2023[98]).
- Polaris estimates growth of the global generative-AI market at a compound annual rate of 34.2 percent, from USD 10.6 billion in 2022 to USD 200.7 billion by 2032 (Polaris, 2023[15]).

At present, generative AI is at an early developmental stage, requiring large investments in R&D and a skilled but scarce workforce to take it to the next stage of maturity. Further growth is expected to come from audio synthesis, data pre-processing, image compression, noise reduction from visual data, medical imaging, and image classification, especially in healthcare (Polaris, 2023[15]).

Application areas include chip and parts design, material sciences, and entertainment.

Gartner, a market research firm, lists other drivers of growth from applying generative AI to chip design, generative design of parts used by industries such as automotive, aerospace, and defence, and to material sciences (Burke and Wiles, 2023[99]). Gartner notes that start-ups building a business on generative AI have received more than USD 1.7 billion in funding over the last three years.

The media and entertainment sector (including advertising) accounts for the largest revenue share from generative AI so far (Polaris, 2023[15]). Companies with a competitive edge in generative AI include well-known, large technology companies, enterprise software providers, AI companies in niche sectors (e.g., legal contract automation, video creation, synthetic data generation, and the arts), and companies providing AI compute such as semiconductors and supercomputing infrastructure, crucial to leveraging generative AI's data-rich environments.

Potential future concerns and risks

Generative AI is expected to exacerbate existing issues associated with AI.

Generative AI can exacerbate issues already on the radar for the OECD and governments and introduce new risks and safety concerns from the race to release novel AI systems and their technological underpinnings. Near-term issues, often rooted in present-day opportunities and challenges, which policy makers should consider due to their urgency and potential for impact include, but are not limited to:

- labour-market impacts, including job displacement, changing skills needs, labour-market inclusiveness, and promoting trustworthy use of AI in the workplace
- information pollution – including the reduced quality of generative AI outputs due to exponential growth in AI-generated content ingested as training data by other AI systems in a vicious cycle – and the consequent decreasing informational relevance of the Internet (Martínez et al., 2023[100])
- AI coding assistants enabling automated cyber-security attacks (Weidinger et al., 2022[29])
- generative AI's role in mass surveillance and censorship (Weidinger et al., 2022[29])

- overreliance and dependency on generative AI systems (Weidinger et al., 2022[29]) (OpenAI, 2023[36])
- copyright issues for new creations and from training on copyrighted works
- academic and creative dishonesty, such as plagiarism (Ka Yuk Chan, 2023[101])
- concentration of AI resources (data, hardware, talent) among few multinational tech companies and governments (Lorenz and Saslo, 2019[102]); (Chawla et al., 2023[3])
- disparate access to generative AI across societies, countries, and world regions
- the need for stronger efforts to curate diverse, high-quality datasets (Bender et al., 2021[58]); (Bender and Koller, 2020[45])
- mis- and disinformation, hate speech, bias, and discrimination by increasingly powerful and realistic generative AI outputs
- generative AI's ecological footprint and natural resources consumption from the tremendous amounts of computing power required for deep learning (Stoken-Walker, 2023[103]).

Risks from emerging model behaviours are also critical to address.

For many years, academic and applied researchers and civil society actors have been steering AI models to align with human values to address a range of potential societal risks (Weidinger et al., 2022[29]); (OpenAI, 2023[36]); (Chan et al., 2023[104]). More recent AI safety research raises issues specifically around generative AI models exhibiting unforeseen "emergent behaviours", such as increased agency, power-seeking, and reward-hacking. Though as mentioned earlier, there is debate over the extent to which these emergent abilities are real versus a "mirage" (Schaeffer, Miranda and Koyejo, 2023[13]).

Researchers identify four characteristics that intensify the agency of algorithmic systems (Chan et al., 2023[104]):

- **Under-specification** – the degree to which the algorithmic system can accomplish a goal provided by operators or designers, without a concrete specification of how the goal is to be accomplished[25]
- **Directness of impact** – the degree to which the algorithmic system's actions affect the world without mediation or intervention by a human (i.e., without a human in the loop)
- **Goal-directedness** – the degree to which the system is designed/trained to achieve a particular quantifiable objective
- **Long-term planning** – the degree to which the algorithmic system is designed/trained to make decisions that are temporally dependent upon one another to achieve a goal and/or make predictions over a long time horizon

Combining these factors can increase agency further. Two major harms that can arise from increased agency of algorithmic systems:

- **Systemic, delayed harms** – non-immediate harms that can be "destructive, long-lasting, and hard to fix", such as social-media recommender systems based on reinforcement-learning. Such algorithms optimise for metrics that can "change or manipulate user's internal states (e.g. preferences, beliefs, psychology)" (Chan et al., 2023[104]).
- **Collective disempowerment** – the perceived danger that model capabilities will perform increasingly important functions in society, taking power away from humans. This could take the form of gradually ceding decision-making to generative AI systems. Its second impact is intensifying concentrations of power and the ability to reap the benefits of AI – already a concern.

AI safety researchers are explicitly looking into another emerging behaviour of concern to the alignment between AI objectives and human preferences: power-seeking, in which goals that provoke power-seeking are reinforced during training and pursued more directly and with novel strategies during deployment,

posing new and potentially severe threats to society (Turner et al., 2019[105]); (Turner and Tadepalli, 2022[106]); (Krakovna and Kramar, 2023[107]); (OpenAI, 2023[36]).

Machine-Learning (ML) systems demonstrate two emergent behaviours that could be catalysed by growing generative AI model capabilities. In *reward hacking*, a model finds unforeseen, and potentially harmful ways of achieving a goal while exploiting the reward signal (Skalse, Howe and Krueger, 2022[108]). In pursuing *instrumental goals*, a model seeks strategies to attain sub-objectives that help it reach an envisaged goal, which might go against the intent of the developers and envisaged goal (Chan et al., 2023[104]). Early evidence shows this can happen even without explicit instructions by model operators or designers. For example, to solve a CAPTCHA code during initial safety testing, ChatGPT misrepresented itself as a vision-impaired human and hired a gig economy worker to solve the CAPTCHA for it. Researchers find that models trained with reinforcement learning from human feedback (RLHF) are more likely to exhibit behaviours such as persuading developers to not shut off the system, pretending to be human, and seeking resource acquisition, such as accruing wealth (Perez et al., 2022[68]); (Chan et al., 2023[104]).

There is growing debate on the path(s) and timeline towards artificial general intelligence

The increasing capabilities of generative AI models have prompted reflection in the media and among AI researchers about whether these models would lead to artificial general intelligence (AGI). Some AI researchers and tech experts did not expect the latest capabilities of generative AI and its possible trajectories (Dardaman and Gupta, 2023[109]). Researchers at Microsoft put forward that GPT-4 "could reasonably be viewed as an early (yet still incomplete) version of an [AGI] system" (Bubeck et al., 2023[110]).

As touched on earlier in this paper, increased agentic model behaviour as a pathway to AGI was demonstrated by researchers at Stanford University and Google Research who created an environment in which 25 generative AI agents interacted over two days and displayed human-like behaviour such as reasoning about their research careers or planning about attending social events (Park et al., 2023[11]). The experiment combined LLMs with interactive agents, which according to the authors, allowed studying human behaviour through increasingly plausible simulations. Commentators were quick to point to possible AGI behaviour among the generative AI agents in the experiment.[26] Other research findings discussed above, such as persuading developers to not shut off the system, relate to often-discussed technical and philosophical concerns of control, such as AI systems' refusal to be shut off, which is beyond the scope of this paper but is being considered in other OECD workstreams (Russell, 2019[111]).

These findings, coupled with research findings—such as the scoring of GPT-4 within the 90th percentile at the Bar Exam (OpenAI, 2023[36]), and finding this model to having reached theory of mind-like cognition levels of a nine-year-old (Kosinski, 2023[112]), help to drive some arguments that that generative AI is moving towards AGI.[27]

The potential benefits and risks from AGI deserve attention because of their potentially broad societal and global impacts. Likewise, narrower generative AI systems deserve focus due to potentially imminent impacts that could be just as significant as those of AGI. Governments should consider the positive and negative implications of both, leveraging strategic foresight and inclusive, long-term policy-making tools.

Risk mitigation measures

Future risks of generative AI could demand solutions on a larger, more systemic scale. These include regulation, ethics frameworks, technical AI standardisation, audits, model release, and access strategies, among others. These and other measures are the topic of a workstream under the G7 Presidency, the results of which will be forthcoming.

4 Conclusion

Generative AI models that generate text, image, video, and audio (e.g., music, speech) content are advancing at breakneck speed. This poses endless possibilities, demonstrated across a growing array of domains. However, the technology also poses numerous challenges and risks to individuals, companies, economies, societies, and policymaking around the globe, ranging from near-term labour-market disruption and disinformation to potential long-term challenges in controlling machine actions. The future trajectories of generative AI are difficult to predict, but governments must explore them to have a hand in shaping them.

Technological development of generative AI is in its nascent stage, with first-movers such as established tech players like Microsoft, Google and Meta, and private research labs like OpenAI, Midjourney, and Stability.AI. These firms are pursuing multiple strategies to capitalise on generative AI and, to some extent, mitigate its downsides.

Public discussion about generative AI is less than a year old. With technology companies bringing generative AI applications to market, policy makers around the globe are grappling with its implications. Applied and academic researchers are engaged in a fierce debate about how to handle generative AI, from mitigation measures in model design and development, through market launch and beyond.

The path ahead is unclear and replete with differing perspectives. One extreme argues for a moratorium on experiments with generative AI more advanced than GPT-4 (Future of Life Institute, 2023[113]) while the other believes that the supposed existential risks of AI are overhyped (LeCun and Ng, 2023[114]).[28] Others—perhaps most—fall somewhere in between. Regardless of ideological stance on these issues, there is an urgent need for further research to prepare for different possible generative-AI future scenarios. Given the great uncertainty and potentially large impact the technology could have at both micro and macro levels, policy makers must remain informed and prepared to take appropriate action through forward-looking AI policies.

The OECD intends for this paper to serve as a steppingstone to help governments make progress in this area. The OECD.AI Policy Observatory and its new OECD Expert Group on AI Futures[29] will serve alongside other relevant bodies as a forum for dialogue on these topics, generating insights and actionable recommendations for governments. Complementary work is also underway through other OECD initiatives, such as work conducted under the Employment, Labour and Social Affairs Committee, the OECD DIS/MIS Resource Hub[30] and the horizontal OECD Going Digital initiative.[31]

References

Abbott, R. and E. Rothman (2022), "Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence", https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html. [7]

Abdelnabi, S. and M. Fritz (2021), *Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding*, https://arxiv.org/abs/2009.03015. [51]

Abid, A., M. Farooqi and J. Zou (2021), "Persistent Anti-Muslim Bias in Large Language Models", http://arxiv.org/abs/2101.05783. [59]

Ada Lovelace Institute (2023), *Explainer: What is a foundation model?*, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer. [4]

Albergotti, R. (2023), "The Secret History of Elon Musk, Sam Altman, and OpenAI", *Semafor*, https://www.semafor.com/article/03/24/2023/the-secret-history-of-elon-musk-sam-altman-and-openai. [93]

Allynn, B. (2022), "Deepfake Video of Zelenskyy Could Be 'tip of the Iceberg' in Info War, Experts Warn", https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia. [43]

Baidoo-Anu, D. and A. Ansah (2023), "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning", https://ssrn.com/abstract=4337484. [21]

Bender, E. et al. (2021), "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23*, Vol. Association for Computing Machinery, Inc., https://doi.org/10.1145/3442188.3445922. [58]

Bender, E. and A. Koller (2020), "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data", https://www.nytimes.com/2018/11/18/technology/artific. [45]

Boháček, M. and H. Farid (2022), "Protecting President Zelenskyy against Deep Fakes", http://arxiv.org/abs/2206.12043. [44]

Bommarito, M. and D. Katz. (2022), "GPT Takes the Bar Exam", *Cornell University*, https://arxiv.org/abs/2212.14402. [115]

Bommasani, R. et al. (2021), "On the Opportunities and Risks of Foundation Models", http://arxiv.org/abs/2108.07258. [23]

Borgeaud, S. et al. (2021), "Improving Language Models by Retrieving from Trillions of Tokens", http://arxiv.org/abs/2112.04426. [49]

Brown, T. et al. (2020), "Language Models Are Few-Shot Learners", http://arxiv.org/abs/2005.14165. [91]

Brynjolfsson, E., D. Li and L. Raymond (2023), "Generative AI at Work", *Nber Workng Paper Series*, https://www.nber.org/system/files/working_papers/w31161/w31161.pdf. [87]

Bubeck, S. et al. (2023), "Sparks of Artificial General Intelligence: Early Experiments with GPT-4", http://arxiv.org/abs/2303.12712. [110]

Buchanan, B. et al. (2021), "Truth, Lies, and Automation How Language Models Could Change Disinformation", https://cset.georgetown.edu/publication/truth-lies-and-automation/. [41]

Burke, B. and J. Wiles (2023), "Beyond ChatGPT: The Future of Generative AI for Enterprises", *Gartner*, https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises. [99]

Chaix, B. et al. (2019), "When Chatbots Meet Patients: One-Year Prospective Study of Conversations between Patients with Breast Cancer and a Chatbot", *JMIR Cancer 5 (1)*, https://doi.org/10.2196/12856. [25]

Chan, A. et al. (2023), "Harms from Increasingly Agentic Algorithmic Systems", *IEEE Computer Society.*, Vol. Vol. 2022-March, https://arxiv.org/abs/2302.10329. [104]

Chawla, S. et al. (2023), *Ten Years after ImageNet: A 360° Perspective on Artificial Intelligence*, Royal Society Open Science 10 (3), https://arxiv.org/abs/2210.01797. [3]

Christiano, P. (2023), *Thoughts on the impact of RLHF research*, https://www.alignmentforum.org/posts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research. [70]

Coldewey, D. and F. Lardinois (2023), *AI is eating itself: Bing's AI quotes COVID disinfo sourced from ChatGPT*, https://techcrunch.com/2023/02/08/ai-is-eating-itself-bings-ai-quotes-covid-disinfo-sourced-from-chatgpt. [39]

Conway, A. (2023), *Bing Chat: What is it, and how does it work?*, https://www.xda-developers.com/bing-chat. [10]

Craig, C. (2021), "The AI-Copyright Challenge: Tech-Neutrality, Authorship, and the Public Interest", https://ssrn.com/abstract=4014811. [79]

Dardaman, E. and A. Gupta (2023), "Bing's Threats Are a Warning Shot", https://abhishek-gupta.ca/aci/blog/bings-threats-are-a-warning-shot. [109]

Dearing, J. (2023), "All Midjourney Versions Compared [V1-V5 Evolution]", https://aituts.com/midjourney-versions/. [96]

Deltorn, J. and A. Macrez (2018), "Authorship in the Age of Machine Learning and Artificial Intelligence", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3261329. [81]

Demner-Fushman, D., Y. Mrabet and A. Abacha (2020), "Consumer Health Information and Question Answering: Helping Consumers Find Answers to Their Health-Related Information Needs", *Journal of the American Medical Informatics Association 27 (2): 194–201*, https://doi.org/10.1093/jamia/ocz152. [24]

Denton, E. et al. (2020), "Bringing the People Back In: Contesting Benchmark Machine Learning Datasets", http://arxiv.org/abs/2007.07399. [65]

Dickson, B. (2023), *How open-source LLMs are challenging OpenAI, Google, and Microsoft*, https://bdtechtalks.com/2023/05/08/open-source-llms-moats/. [5]

Dodge, J. et al. (2021), "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus", http://arxiv.org/abs/2104.08758. [63]

Dohmke, T. (2022), "GitHub Copilot Is Generally Available to All Developers", https://github.blog/2022-06-21-github-copilot-is-generally-available-to-all-developers/. [16]

Eloundou, T. et al. (2023), "GPTs Are GPTs: An Early Look at the Labour Market Impact Potential of Large Language Models", http://arxiv.org/abs/2303.10130. [84]

Felten, E., M. Raj and R. Seamans (2023), "How Will Language Modelers like ChatGPT Affect Occupations and Industries?", http://arxiv.org/abs/2303.01157. [85]

Fraser, K., S. Kiritchenko and I. Nejadgholi (2023), "A Friendly Face: Do Text-to-Image Systems Rely on Stereotypes When the Input Is Under-Specified?", http://arxiv.org/abs/2302.07159. [62]

Future of Life Institute (2023), *Pause Giant AI Experiments: An Open Letter*, https://futureoflife.org/open-letter/pause-giant-ai-experiments. [113]

Galaz, V. et al. (2023), *AI Could Create a Perfect Storm of Climate Misinformation*, Stockholm Reslience Centre, https://www.stockholmresilience.org/download/18.889aab4188bda3f44912a32/1687863825612/SRC_Climate%20misinformation%20brief_A4_.pdf. [38]

Ganguli, D. et al. (2022), "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviours, and Lessons Learned", http://arxiv.org/abs/2209.07858. [52]

Gebru, T. et al. (2018), "Datasheets for Datasets", http://arxiv.org/abs/1803.09010. [64]

Goldman Sachs (2023), "Generative AI could raise global GDP by 7%", https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html. [97]

Goldstein, J. et al. (2023), "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations", http://arxiv.org/abs/2301.04246. [40]

Gómez, E. et al. (2018), *Deep Learning for Singing Processing: Achievements, Challenges and Impact on Singers and Listeners*, https://arxiv.org/abs/1807.03046. [20]

Goodfellow, I., Y. Bengio and A. Courville (2016), *Deep Learning*, MIT Press, https://www.deeplearningbook.org. [2]

Goodfellow, I. et al. (2014), "Generative Adversarial Nets", *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, https://papers.nips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf. [9]

Groh, M. et al. (2022), "Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds", *Proceedings of the National Academy of Sciences*, Vol. 119/1, https://arxiv.org/abs/2105.06496. [53]

Henderson, P. et al. (2023), *Foundation Models and Fair Use*, https://arxiv.org/abs/2303.15715. [77]

Heusel, M. et al. (2018), "GANs Trained by a Two Time-Scale Update Rule", *31st Conference on Neural Information Processing Systems*, https://arxiv.org/pdf/1706.08500.pdf. [116]

House of Representatives (1976), "Historical and Revision Notes House Report No. 94-1476", https://www.govinfo.gov/content/pkg/USCODE-2010-title17/pdf/USCODE-2010-title17-chap1-sec107.pdf. [74]

Hu, K. (2023), "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note", *Reuters*, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. [8]

Hurler, K. (2023), *Chat-GPT Pretended to Be Blind and Tricked a Human Into Solving a CAPTCHA*, https://gizmodo.com/gpt4-open-ai-chatbot-task-rabbit-chatgpt-1850227471. [14]

Hutchinson, B. et al. (2021), "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure", *Association for Computing Machinery, Inc.* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 560–75, https://doi.org/10.1145/3442188.3445918. [66]

Infopaq International (C-5/08) (2009), "European Court of Justice [C-5/08] ECLI:EU:C:2009:465", https://curia.europa.eu/juris/liste.jsf?num=c-5/08. [80]

Ingle, P. (2023), *Top Artificial Intelligence (AI) Tools That Can Generate Code To Help Programmers*, https://www.marktechpost.com/2023/07/11/top-artificial-intelligence-ai-tools-that-can-generate-code-to-help-programmers. [18]

Insilico Medicine (2023), *Insilico Medicine receives IND approval for novel AI-designed USP1 inhibitor for cancer*, https://www.eurekalert.org/news-releases/990417. [26]

Ka Yuk Chan, C. (2023), *Is AI Changing the Rules of Academic Misconduct? An In-depth Look at Students' Perceptions of 'AI-giarism'*, https://arxiv.org/abs/2306.03358. [101]

Kaplan, J. et al. (2020), "Scaling Laws for Neural Language Models", http://arxiv.org/abs/2001.08361. [92]

Kirchenbauer, J. et al. (2023), *A Watermark for Large Language Models*, https://arxiv.org/abs/2301.10226. [50]

Kosinski, M. (2023), "Theory of Mind May Have Spontaneously Emerged in Large Language Models", https://osf.io/csdhb. [112]

Krakovna, V. and J. Kramar (2023), *Power-Seeking Can Be Probable and Predictive for Trained Agents*, https://arxiv.org/abs/2304.06528. [107]

Kreps, S., R. Mccain and M. Brundage. (2022), "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation", *Journal of Experimental Political Science 9 (1)*, pp. 104–17, https://doi.org/10.7910/DVN/1XVYU3. [28]

Krishna, K. et al. (2023), "Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense", http://arxiv.org/abs/2303.13408. [57]

Lambert, N. et al. (2022), *Illustrating Reinforcement Learning from Human Feedback (RLHF)*, https://huggingface.co/blog/rlhf. [71]

Le, B. et al. (2023), "Why Do Deepfake Detectors Fail?", http://arxiv.org/abs/2302.13156. [55]

LeCun, Y. (2022), "A Path Towards Autonomous Machine Intelligence", https://openreview.net/pdf?id=BZ5a1r-kVsf. [35]

LeCun, Y. and A. Ng (2023), *Yann LeCun and Andrew Ng: Why the 6-month AI Pause is a Bad Idea*, https://www.youtube.com/live/BY9KV8uCtj4. [114]

Lin, S., J. Hilton and A. Evans (2022), "ruthfulQA: Measuring How Models Mimic Human Falsehoods", https://github.com/sylinrl/TruthfulQA. [46]

Longoni, C. et al. (2022), "News from Generative Artificial Intelligence Is Believed Less", *In ACM International Conference Proceeding Series, 97–106*, Vol. Association for Computing Machinery, https://dl.acm.org/doi/abs/10.1145/3531146.3533077. [32]

Lorenz, P. (2022), "Analyzing Global AI Dependencies through Intellectual Property Rights -- Understanding Trade Secrets, Patents, and Copyrights for Artificial Intelligence", https://www.stiftung-nv.de/sites/default/files/snv_analyzing_global_ai_dependencies_through_intellectual_property_rights.pdf. [75]

Lorenz, P. and K. Saslo (2019), *Demystifying AI and AI Companies – What Foreign Policy Makers Need to Know About the Global AI Industry*, https://www.stiftung-nv.de/de/publikation/demystifying-ai-ai-companies-what-foreign-policy-makers-need-know-about-global-ai. [102]

Luccioni, A. et al. (2023), "Stable Bias: Analyzing Societal Representations in Diffusion Models", http://arxiv.org/abs/2303.11408. [60]

Lucy, L. and D. Bamman (2021), "Gender and Representation Bias in GPT-3 Generated Stories", https://github.com/lucy3/gpt3_gender. [61]

Maerten, A. and A. Soydaner (2023), "From Paintbrush to Pixel: A Review of Deep Neural Networks in AI-Generated Art", http://arxiv.org/abs/2302.10913. [95]

Marcus, G. (2020), "he Next Decade in AI: Four Steps Towards Robust Artificial Intelligence", http://arxiv.org/abs/2002.06177. [94]

Markov, T. et al. (2022), "A Holistic Approach to Undesired Content Detection in the Real World", http://arxiv.org/abs/2208.03274. [69]

Martínez, G. et al. (2023), "Combining Generative Artificial Intelligence (AI) and the Internet: Heading towards Evolution or Degradation?", http://arxiv.org/abs/2303.01255. [100]

McKinsey (2023), "The economic potential of generative AI: The next productivity frontier", https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier. [98]

Menick, J. et al. (2022), "Teaching Language Models to Support Answers with Verified Quotes", http://arxiv.org/abs/2203.11147. [48]

Morandini, S. et al. (2023), "The Impact of Artificial Intelligence on Workers' Skills: Upskilling and Reskilling in Organisations", *Informing Science: The International Journal of an Emerging Transdiscipline 26: 039–068*, https://doi.org/10.28945/5078. [90]

Murray, M. (2023), "Generative and AI Authored Artworks and Copyright Law", *Hastings Communications and Entertainment Law Journal*, Vol. 45, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4152484. [78]

Nakano, R. et al. (2021), "WebGPT: Browser-Assisted Question-Answering with Human Feedback", http://arxiv.org/abs/2112.09332. [47]

Nightingale, S. and H. Farid. (2022), "AI-Synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 119/8, https://doi.org/10.1073/pnas.2120481119. [6]

Nijkamp, E. et al. (2023), *CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis*, https://arxiv.org/abs/2203.13474. [17]

Noy, S. et al. (2023), "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence", https://www.science.org/doi/10.1126/science.adh2586. [86]

OECD (2023), "AI language models: Technological, socio-economic and policy considerations", *OECD Digital Economy Papers*, No. 352, OECD Publishing, Paris, https://cdn.openai.com/papers/gpt-4-system-card.pdf. [1]

OECD (2023), *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*, OECD Publishing, https://doi.org/10.1787/08785bba-en. [83]

OECD (2022), *Building Trust and Reinforcing Democracy*, OECD Publishing, https://doi.org/10.1787/76972a4a-en. [31]

Ognyanova, K. et al. (2020), "Misinformation in Action: Fake News Exposure Is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is in Power", *Harvard Kennedy School Misinformation Review*, https://doi.org/10.37016/mr-2020-024. [30]

OpenAI (2023), "GPT-4 System Card OpenAI", https://cdn.openai.com/papers/gpt-4-system-card.pdf. [36]

Park, J. et al. (2023), "Generative Agents: Interactive Simulacra of Human Behaviour", http://arxiv.org/abs/2304.03442. [11]

Passi, S. and M. Vorvoreanu (2022), "Overreliance on AI: Literature Review", https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf. [37]

Peng, S. et al. (2023), "he Impact of AI on Developer Productivity: Evidence from GitHub Copilot", http://arxiv.org/abs/2302.06590. [88]

Perez, E. et al. (2022), "Discovering Language Model Behaviours with Model-Written Evaluations", http://arxiv.org/abs/2212.09251. [68]

Pettinato Oltz, T. (2023), "ChatGPT, Professor of Law", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4347630. [22]

Polaris (2023), "Generative AI Market Share, Size, Trends, Industry Analysis Report, By Component (Software and Services); By Technology; By End-Use; By Region; Segment Forecast, 2023 - 2032", https://www.polarismarketresearch.com/industry-analysis/generative-ai-market. [15]

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking. [111]

Sadasivan, V. et al. (2023), "Can AI-Generated Text be Reliably Detected?", *Cornell University*, https://arxiv.org/abs/2303.11156. [54]

Sadasivan, V. et al. (2023), "Can AI-Generated Text Be Reliably Detected?", http://arxiv.org/abs/2303.11156. [56]

Schaeffer, R., B. Miranda and S. Koyejo (2023), "Are Emergent Abilities of Large Language Models a Mirage?", *Cornell University*, https://arxiv.org/abs/2304.15004. [13]

Shah, D. (2023), *RLHF (Reinforcement Learning From Human Feedback): Overview + Tutorial*, https://www.v7labs.com/blog/rlhf-reinforcement-learning-from-human-feedback. [72]

Skalse, J., N. Howe and D. Krueger (2022), "Defining and Characterizing Reward Hacking", https://arxiv.org/abs/2209.13085. [108]

Sofer, C. et al. (2015), "What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness", *Psychological Science 26 (1): 39–47*, https://doi.org/10.1177/0956797614554955. [33]

Sriram, A. and C. Mehta (2023), "OpenAI Tech Gives Microsoft's Bing a Boost in Search Battle with Google", *Reuters*, https://www.reuters.com/technology/openai-tech-gives-microsofts-bing-boost-search-battle-with-google-2023-03-22/. [27]

Stoken-Walker, C. (2023), *https://www.wired.co.uk/article/the-generative-ai-search-race-has-a-dirty-secret*, https://www.wired.co.uk/article/the-generative-ai-search-race-has-a-dirty-secret. [103]

Turner, A. et al. (2019), "Optimal Policies Tend to Seek Power", http://arxiv.org/abs/1912.01683. [105]

Turner, A. and P. Tadepalli (2022), "Parametrically Retargetable Decision-Makers Tend To Seek Power", http://arxiv.org/abs/2206.13477. [106]

U.S. Copyright Office (2023), "https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance", *Statement of Policy*, https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence. [82]

Vaswani, A. et al. (2023), "Attention Is All You Need (v7)", *Proceedings of the 31st Conference on Neural Information Processing Systems*, https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. [117]

Wei et al. (2022), "Emergent Abilities of Large Language Models", *Transactions on Machine Learning Research*, https://openreview.net/pdf?id=yzkSU5zdwD. [12]

Weidinger, L. et al. (2022), "Taxonomy of Risks Posed by Language Models", *ACM International Conference Proceeding Series*, Vol. Association for Computing Machinery, pp. 214–29, https://doi.org/10.1145/3531146.3533088. [29]

Weidinger, L. et al. (2022), "Taxonomy of Risks Posed by Language Models", *In ACM International Conference Proceeding Series, 214–29*, https://doi.org/10.1145/3531146.3533088. [34]

WIPO (2020), "WIPO Conversations on Intellectual Property and Artificial Intelligence", https://roadtobern.sw. [73]

Yang, L., S. Chou and Y. Yang (2017), "MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation", http://arxiv.org/abs/1703.10847. [19]

Zakrzewski, C. (2023), *Biden administration is trying to figure out how to audit AI*, https://www.washingtonpost.com/technology/2023/04/11/biden-commerce-department-ai-rules (accessed on 12 July 2023). [67]

Zarifhonarvar, A. (2023), "Economics of ChatGPT: A Labour Market View on the Occupational Impact of Artificial Intelligence", https://ssrn.com/abstract=4350925. [89]

Zellers, R. et al. (2019), "Defending Against Neural Fake News", http://arxiv.org/abs/1905.12616. [42]

Zirpoli, C. (2023), "Generative Artificial Intelligence and Copyright Law", https://crsreports.congress.gov. [76]

Notes

[1] The underlying neural network architecture enabling text-to-image capabilities are diffusion models for image synthesis combined with transformers relevant for text inputs, or Generative Adversarial Networks (GANs). See https://www.sabrepc.com/blog/Deep-Learning-and-AI/gans-vs-diffusion-models.

[2] The G7 is an intergovernmental forum of large economies comprising Canada, the European Union, France, Germany, Italy, Japan, the United Kingdom, and the United States. Its May 2023 statement can be found at https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/20/g7-hiroshima-leaders-communique.

[3] For an overview on FOSS for AI, see Chiradeep BasuMallick (2021). "Top 10 Open Source Artificial Intelligence Software in 2021". https://www.spiceworks.com/tech/innovation/articles/top-open-source-artificial-intelligence-software.

[4] See https://stability.ai/stablediffusion and https://ai.meta.com/llama, respectively.

[5] Today's most widely used model for text-to-text generation is generative pre-trained transformers (GPTs), invented by Google in 2017 (Vaswani et al., 2023[117]).

[6] Matt Novak. "That Viral Image Of Pope Francis Wearing A White Puffer Coat Is Totally Fake". Forbes. March 26, 2023. https://www.forbes.com/sites/mattnovak/2023/03/26/that-viral-image-of-pope-francis-wearing-a-white-puffer-coat-is-totally-fake/?sh=7d8c87b71c6c.

[7] See https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/20/g7-hiroshima-leaders-communique.

[8] For details on the requirements, see Dan Cooper et al. "Preview into the European Parliament's Position on the EU's AI Act Proposal". March 28, 2023. Covington. https://www.insideprivacy.com/artificial-intelligence/a-preview-into-the-european-parliaments-position-on-the-eus-ai-act-proposal.

[9] https://www.khanacademy.org/khan-labs.

[10] Insilico Medicine receives IND approval for novel AI-designed USP1 inhibitor for cancer, https://www.eurekalert.org/news-releases/990417.

[11] https://www.bing.com/new.

[12] See the OECD dis- and misinformation resource hub for detail: https://www.oecd.org/stories/dis-misinformation-hub.

[13] Overreliance is defined as "users accepting incorrect AI recommendations"", as discussed in "Overreliance on AI: Literature review". Passi and Vorvoreanu 2022. https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf.

[14] These biases may be rooted in the training data (problematic relationships between captions and images), model design (aesthetic bias towards youth and femininity found in models steered towards producing AI artwork), or the proactive vs. hands-off vs. design choices of the AI engineers, which either try to actively reduce biases or shift responsibility for appropriate use to the user. (Fraser, Kiritchenko and Nejadgholi, 2023[62]).

[15] See, for example, "external red teaming" in "DALL·E 2 Preview - Risks and Limitations". OpenAI. 2022. https://github.com/openai/dalle-2-preview/blob/main/system-card.md.

[16] See https://huggingface.co/blog/rlhf for an overview of RLHF.

[17] These are: "(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.", see: House of Representatives. 1976. Copyright Law Revision. https://www.govinfo.gov/content/pkg/USCODE-2010-title17/pdf/USCODE-2010-title17-chap1-sec107.pdf. p. 25.

[18] As an example, see https://stablediffusionlitigation.com/pdf/00201/1-1-stable-diffusion-complaint.pdf.

[19] Some harmonisation is nevertheless provided by international treaties such the Berne Convention for the Protection of Literary and Artistic Works: https://www.wipo.int/treaties/en/ip/berne/; https://www.wipo.int/wipolex/en/text/283693.

[20] See the UK Copyright, Designs and Patents Act 1988, chapter 48, Section 9(3) Authorship of work: "In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken." https://www.legislation.gov.uk/ukpga/1988/48/section/9/enacted. See also New Zealand Copyright Act 1994 Section 5(2)(a) Meaning of Author: "For the purposes of subsection (1), the person who creates a work shall be taken to be,— (a) in the case of a literary, dramatic, musical, or artistic work that is computer-generated, the person by whom the arrangements necessary for the creation of the work are undertaken (…)" https://www.legislation.govt.nz/act/public/1994/0143/latest/DLM345899.html.

[21] https://copyright.gov/ai

[22] (Bommarito and Katz., 2022[115]) describe the preparation for the Bar Exam in the US: "Nearly all jurisdictions in the United States require a professional license exam, commonly referred to as "the Bar Exam," as a precondition for law practice. To even sit for the exam, most jurisdictions require that an applicant completes at least seven years of post-secondary education, including three years at an accredited law school. In addition, most test-takers also undergo weeks to months of further, exam-specific preparation. Despite this significant investment of time and capital, approximately one in five test-takers still score under the rate required to pass the exam on their first try."

[23] With models such as for example Chinchilla (DeepMind) Hoffmann et al. 2022. "Training Compute-Optimal Large Language Models". https://arxiv.org/pdf/2203.15556.pdf.; LLaMa (Meta), Touvron et al.

2023. “LLaMA: Open and Efficient Foundation Language Models”. https://arxiv.org/pdf/2302.13971.pdf.; and Alpaca (Stanford University), Taori et al. 2023. “Alpaca: A Strong, Replicable Instruction-Following Model”. https://crfm.stanford.edu/2023/03/13/alpaca.html.

[24] Based on Fréchet Inception Distance (FID) scores, which measure the similarity of generated images to real ones. See (Heusel et al., 2018[116]).

[25] See https://arxiv.org/pdf/2011.03395.pdf for more information.

[26] As discussed for instance in these subreddits: https://www.reddit.com/r/singularity/comments/12ihk72/stanfordgoogles_generative_agents_are_full/; https://www.reddit.com/r/singularity/comments/12hiebn/stanfordgoogle_researchers_just_told_us_how_they/; https://www.reddit.com/r/MachineLearning/comments/12hluz1/r_generative_agents_interactive_simulacra_of/; https://www.reddit.com/r/Futurology/comments/12iemwm/ai_bots_were_given_freedom_in_a_virtual_city_they/.

[27] Theory of mind is “the ability to impute unobservable mental states to others.” (Kosinski, 2023[112]).

[28] In this video, Yann LeCun argues his view on a potential moratorium alongside Andrew Ng, adjunct professor at Stanford University and co-founder of Coursera and deeplearning.ai. “Yann LeCun and Andrew Ng: Why the 6-month AI Pause is a Bad Idea”. https://www.youtube.com/watch?v=BY9KV8uCtj4&t=304s.

[29] See https://oecd.ai/en/network-of-experts/working-group/10847.

[30] https://www.oecd.org/stories/dis-misinformation-hub.

[31] https://www.oecd.org/digital/going-digital-project.