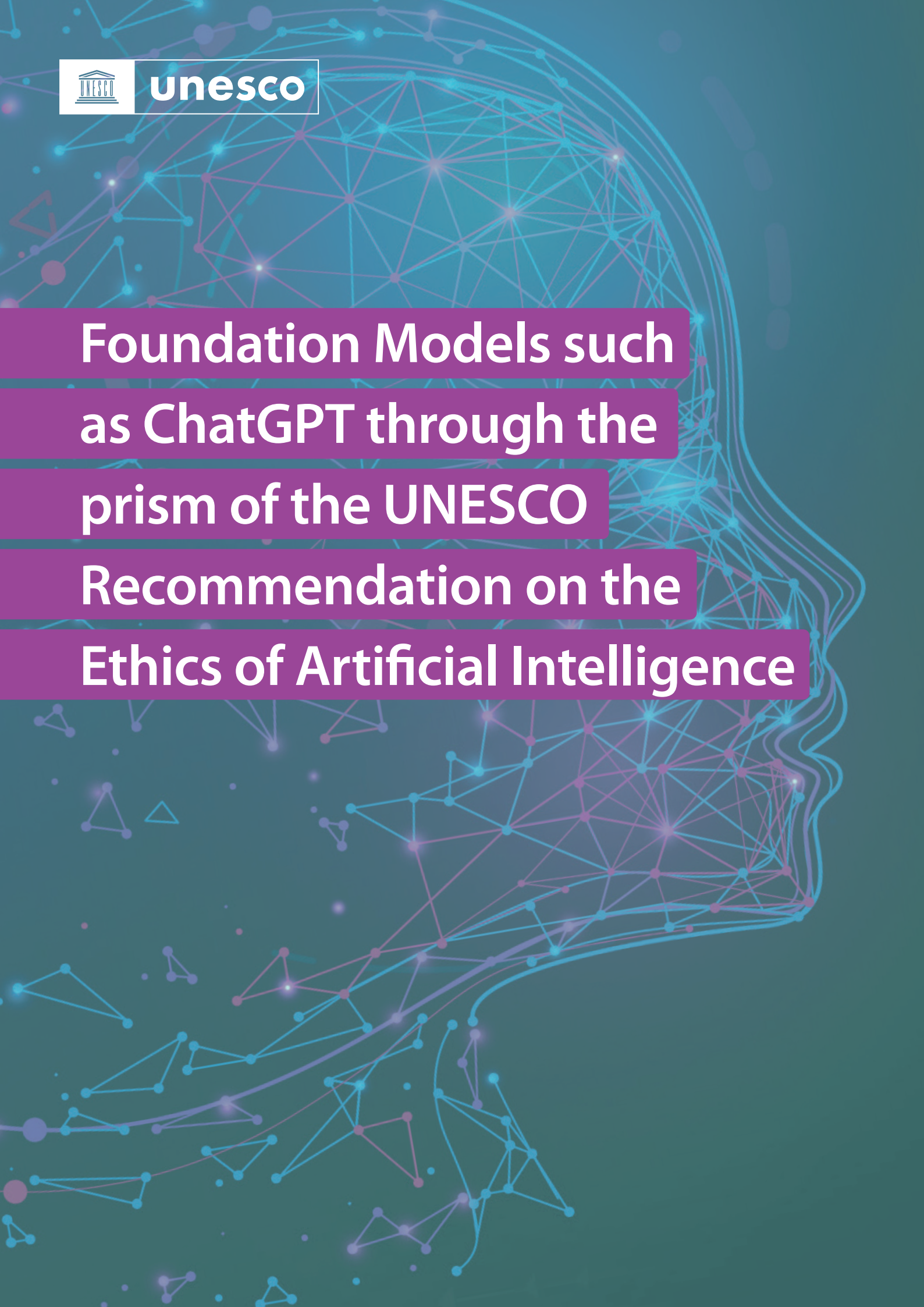

unesco
Foundation Models such as ChatGPT through the prism of the UNESCO Recommendation on the Ethics of Artificial Intelligence

Foundation Models such as ChatGPT through the prism of the UNESCO Recommendation on the Ethics of Artificial Intelligence

Foreword

In recent months, the release of new artificial intelligence (AI) models that can generate text and images based on user prompts has captured global attention. Their speed and depth of deployment means that in just a few months, they have been used by hundreds of millions of people and are rapidly becoming household names. These foundation models – machine learning models trained on huge data sets using immense computing resources – open up many new possibilities for users with potentially transformative implications for how they learn, work, communicate, and find and synthesize information. However, it is already clear that these models could be associated with potential harms on an equally large scale.

Indeed, the well-known risks of AI related to biased and discriminatory outcomes, safety and reliability concerns, and impacts on labour markets and children and youth, among others, have grown significantly in line with the enhanced capacities of LLMs. Preliminary assessments confirm LLMs can deliver misleading, inaccurate, or false information without making this clear to the user (ChatGPT introduced a disclaimer only recently). Their impact on science, research, education, and work is also magnified by the range of tasks the tool can perform. Issues of authorship and intellectual property rights are also paramount, as the platform does not quote its sources and lacks transparency on how it works. This adds to the list of unknowns which augment risks in the human-machine interaction. Generative AI can shape people's minds and thoughts, and therefore also human behaviors.

A recognition of the risks posed by AI systems, and the need to identify and prevent or mitigate them, is what led UNESCO to produce the Recommendation on the Ethics of Artificial Intelligence, which was adopted by all 193 Member States in November 2021. The Recommendation sets out four values premised on the promotion of human rights and human dignity, which are then unpacked in the form of ten principles as well as specific policy recommendations for governments. The aim of the Recommendation is to ensure fair and inclusive outcomes, while enhancing the positive impacts of these technologies. We are already implementing this global standard in a large set of countries and enhancing the capacities of governments and the public at large to deal with AI systems. Addressing the downsides promises to foster innovation and growth by increasing people's trust in using AI. The dichotomy between legislation and innovation that has dominated discussions does not hold, as effective regulatory frameworks provide for certainty and interoperability, allowing companies to flourish. They can also help to level the playing field and enhance competition, benefiting small and medium sized companies.

In this paper, we assess foundation models such as ChatGPT through the lens of the provisions of the UNESCO Recommendation in order to clarify and highlight some of the risks associated with their use, and to suggest a framework with which to address and mitigate them when designing, developing, and deploying AI systems including foundation models.

These AI models are often described as "experimental" by their developers, and it is often only after they have been released to the public that harms start to become apparent, even when these could and should have been anticipated at the design

and development stages. One of the key messages of this discussion paper is that ethical considerations and processes to support them must be built into every stage of the life cycle of such models, in an ex-ante manner to identify and address risks effectively, and to prevent ethics being sidelined while other considerations such as commercial or economic competition prevail.

Many voices are now calling for a review of the way these technologies are developed and launched, signaling the need for stronger governance and oversight capacities. This is an effort that UNESCO has been making for decades in relation to emerging technologies, be they human genome, AI, or neuro technologies, through its ethical mandate. Since 2021, when the AI global standard was adopted, we have been building the tools and support systems for its implementation, and for the review of relevant rules and regulations to ensure good governance of AI without impinging on innovation. The time is ripe to build better rules so that technological developments support our human goals and deliver for the public good.

To advance this work, we are relying on a large group of partners in the public and private sectors as well as civil society to ensure that the Recommendation translates into concrete policies and regulatory insights. With the support of a High-Level Expert Group, representing all regions of the world, we developed the Readiness Assessment, a diagnostic tool to understand where countries stand in their capacities to adopt and govern AI, and the Ethical Impact Assessment to support procurement offices. We are now deploying these tools in a large group of countries with the support of the Japanese Government, the Patrick McGovern Foundation, the European Commission, and the Andina de Fomento Corporation. We have established the *AI Experts without Borders* and *Women4EthicalAI* networks. We are also working with a large set of knowledge institutions and will launch the *Observatory of Ethics of AI* with The Alan Turing Institute. As the private sector produces the largest share of these technologies, we are partnering with Microsoft and Telefonica, who chair our Business Council for the implementation of UNESCO's Recommendation.

ChatGPT and LLMs are creating high expectations of the services they can provide to humanity. These could be significant. However, their widespread use is also highlighting the risks attached to how these technologies are currently being deployed, responding to a frantic technological race between economic actors and countries, instead of serving the public good. To get it right, we need the right oversight and policy frameworks, and this is what UNESCO has been mandated to do by its Member States since 2021. We hope that the concerns that these technologies are raising will help us build more solid governance frameworks to positively impact our economies and societies.

Gabriela Ramos
Assistant Director-General for Social and Human Sciences

Introduction

The release into the public domain and massive growth in the user base of artificial intelligence (AI) foundation models for text, images, and audio is fueling debate about the risks they pose to work, education, scientific research, and democracy, as well as their potential negative impacts on cultural diversity and cross-cultural interactions, among other areas. Foundation models are AI systems that are characterized by the use of very large machine learning models trained on massive unlabeled data sets using considerable compute resources. Examples include large language models (LLMs) such as the GPT series and Bard, and image generator tools such as DALL·E 2 and Stable Diffusion. This discussion paper focuses on a widely used foundation model, ChatGPT, as a case study, but many of the points below are applicable to other LLMs and foundation models more broadly.

Technological development cannot be halted, nor would it be desirable for this to happen. However, an ethical and multistakeholder approach is needed from the start of the AI system project lifecycle to identify, assess, and respond to potential harms, while weighing the benefits not just individually but also for the public good, before it is released to the public on a large scale. This discussion paper leverages the ethical framework provided by the UNESCO Recommendation on the Ethics of Artificial Intelligence and the expertise of the High-Level Expert Group (HLEG) supporting the implementation of the Recommendation to explore ChatGPT in dialogue with emerging critical perspectives on the subject.

The goal of this paper is to demonstrate how the lens of the Recommendation can help identify and clarify key ethical concerns related to AI foundation models such as ChatGPT, and provide the procedural framework to address and mitigate these concerns, including via effective governance models and tools such as ethical impact assessment and complementary approaches such as ethics by design or research ethics committees.

What is ChatGPT, and why does it matter?

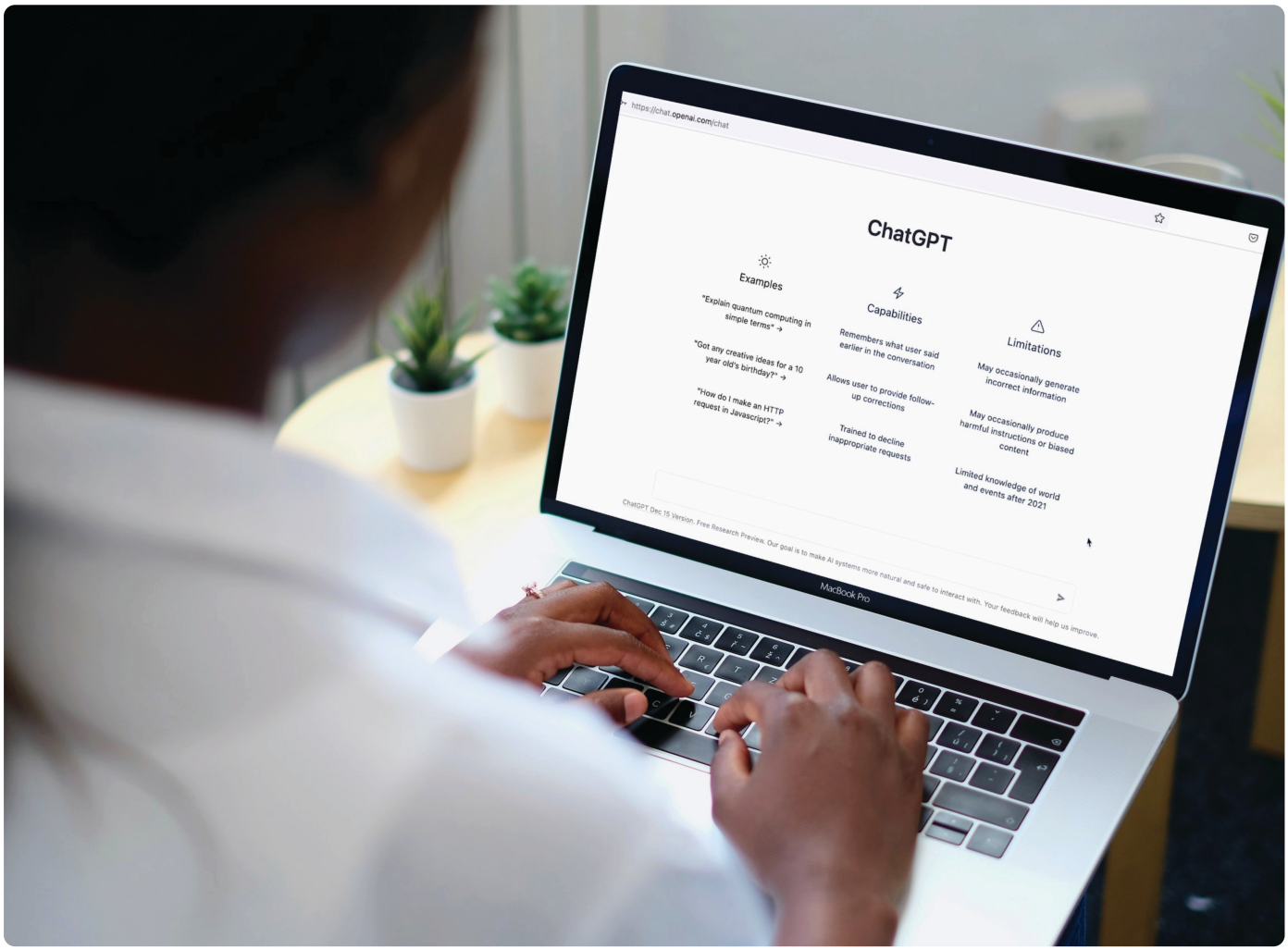
ChatGPT, built on GPT-3.5 (OpenAI, 2022), is considered one of the most sophisticated language models available in the public domain and has been widely utilized globally. ChatGPT is a popular chatbot that can generate coherent and contextually relevant human-like responses to input prompts (OpenAI, 2022). Within five days of its launch in November 2022, ChatGPT gained one million users, outpacing the growth of social media platforms. As of January 2023, it was estimated to have 100 million accounts, with 13 million daily users (Hu, 2023).

Four characteristics of ChatGPT are relevant to the discussion contained in the present paper (Brown et al., 2020).

1. ChatGPT is a generative model trained to perform a wide range of text-based functions, making its applications easily widespread and potentially ubiquitous.
2. ChatGPT relies on deep neural networks that are impossible for humans to understand or explain.
3. Internet crawlers, i.e., software designed to scroll through (or read) and record online content, were heavily employed to gather the massive corpora of data used to train ChatGPT. This will inevitably result in data containing biases, harmful content, or misinformation, for instance, existing on the net.
4. ChatGPT is ahead of its competition in the public domain in terms of the speed of growing its user base (Hu, 2023).

The use of ChatGPT to date has shown its effectiveness in multiple contexts including, but not limited to, translation, writing essays, answering questions in the style of experts, generating computer code, games, songs and creative writing, assisting research, and even obtaining passing grades in medicine and law exams (Cain, 2023; Choi et al., 2023; Dowling & Lucey, 2023; Enoch, 2022; Guo et al., 2023; Imgur, 2022; Jiao et al., 2023). As the scale and scope of its use increases, so does its impact.

Its supporters claim that ChatGPT has the potential to create positive outcomes, while its detractors argue that it poses equally large-scale risks to human rights, societal trust, and inclusion, and increases the likelihood of further concentrating wealth and power. As McQuillan writes of ChatGPT, "the social benefits are still speculative while the harms have been empirically demonstrated" (McQuillan, 2023).



Among its claimed benefits for society at large, some highlight its potential to increase productivity and improve access to a wide range of services for people in economically and socially disadvantaged positions. Also, according to some, its use can improve decision-making and enhance and personalize user experiences, particularly in education and healthcare.

Conversely, among the pernicious effects of ChatGPT and other language models, six dimensions have been highlighted (Zhuo et al., 2023): (1) discrimination and exclusion, resulting from outdated or biased data (2) facilitating the spread of inaccurate information, weakening the verifiability and trustworthiness of the outputs (3) malicious uses, as even the safety and security of the tools as well as their robustness against cyber attack have not been guaranteed before their release (4) human-computer interaction harms, (5) automation, access, and environmental harms, (6) potential infringement of privacy and the protection of data, as it is not clear what measures are in place to protect users' personal data, and there have already been reports of data breaches and the leakage of personal information including names, email addresses, payment addresses, and last four digits of credit card numbers (Powell, 2023). In the next section, we will examine these and other possible adverse effects using the UNESCO Recommendation on the Ethics of Artificial Intelligence as a framework.

The discussion contained in the present discussion paper about the ethical and social implications of ChatGPT extends to most generative AI, i.e. foundation models capable of creating apparently "original" content, including text, graphics, video, voice recordings and other outputs. Since 2022, at least 23 other generative AI models have been launched. Of particular concern is the potential negative effect of AI-generated deepfakes on the reliability and verifiability of information, as well as on public trust in institutions, as such systems can be combined to create new content that is practically impossible to differentiate from originals (Gozalo-Brizuela & Merchan-Garrido, 2023).

ChatGPT analyzed through the lens of the UNESCO Recommendation on the Ethics of Artificial Intelligence

In what follows, this paper examines ChatGPT in relation to the principles and policy areas of the Recommendation. These encompass concerns related to fairness and non-discrimination, inaccurate information, responsibility and accountability, safety and security, privacy and data protection, human and environmental flourishing, and education and research. These principles lead to concrete policy actions to ensure fair, inclusive, and sustainable outcomes from AI developments.

Fairness and non-discrimination

Stereotypical and discriminatory outputs are perhaps the most visible and controversial adverse effects of ChatGPT. These occur mainly due to algorithms replicating the biases contained in the data on which they are trained. Such biases stem from e.g. the disproportional representation of certain populations in the data and the absence of others, including speakers of less-used languages or members of smaller cultural groups (Zhuo et al., 2023). Examples of recent discriminatory results that have made the news are narratives telling racist jokes and associating the word "white" with "superiority"; describing white and Asian men as better scientists; or classifying US and Canadian workers as "senior" and Mexican workers as "junior" (Alba, 2022; Bhadani, 2022; Johnson, 2023). Other generative AI models, like image-creators, are known to propose similar stereotypes, for examples, portraying "lawyers" as white men and "flight attendants" as Asian women (Samuel, 2022).

Gender discrimination in ChatGPT is also a significant concern. Narratives generated by GPT-3 have been shown to reinforce gender stereotypes, depicting female characters as less powerful and defining them by their physical appearance and family roles (Li & Bamman, 2021). This can happen when the system is trained on literary archives in a way that can further perpetuate and reinforce historical stereotypes and prejudices. In relation to religious biases, prejudices are observed in analogies and stories produced by the tool. For example, researchers from Stanford found that Muslims were depicted as terrorists in 23% of the prompt they tested, while Jews were associated with money in 5% (Abid et al., 2021). The potential to discriminate and reinforce traditional biases and prejudices through AI algorithms is a well-known and well-studied phenomenon (Noble, 2018; O'Neil, 2016; Benjamin, 2019). However, a unique feature of foundation models is the widespread overarching coverage that spans all possible domains and topics. While in the past it was possible to examine one specific model and assess whether it was discriminatory or biased against a certain group, as users can now have conversations with those tools about any topic, it is more challenging to foresee and check for or measure biases.

Lack of diversity is also demonstrated in the poor performance of these tools in many languages other than English (Seghier, 2023). Researchers have pointed out differences when addressing queries in English versus other languages and revealed that even if the tools is able to translate data accurately, when it comes to cultural inferences and specific knowledge, the information provided is often based on U.S.-derived perspectives (Walker, 2022).

In relation to the issues mentioned above, the principle on Fairness and Nondiscrimination, outlined in paragraphs 28, 29, and 30 of the Recommendation, underlines the importance of safeguarding fairness and non-discrimination in promoting social justice through AI systems. Specifically, it prioritizes the inclusion of all members of society, emphasizing people with disabilities, women and children, and all marginalized groups, with consideration for their specific needs and language requirements. In addition, the Recommendation highlights the need to address the digital divide and prevent the reinforcement or perpetuation of biases and stereotypes in AI systems.

In addition, in Policy Area 6 on Gender, the Recommendation calls on Member States and other stakeholders including developers to "ensure that the potential of AI systems to advance the achievement of gender equality is realized. They should ensure that these technologies do not exacerbate the already wide gender gaps existing in several fields in the analogue world, and instead eliminate those gaps."

Economy and labour

Foundation models increase concerns about the impact of AI on labour markets, and the speed and depth with which certain jobs will be transformed. For example, it has been estimated that they could have an impact on 80% of the U.S. workforce, affecting approximately 10% of their work tasks (Eloundou et al. 2023). These tools can be used to automate tasks traditionally associated with human functions that

include reasoning, writing, creating graphics, and analyzing data. This challenges the organization of the workforce and the capacity of people to transition to different job profiles. Additionally, not all benefit equally from the productivity growth that AI is supposed to bring. Some people are likely to be automated out of their jobs, and big companies are integrating the tools into their products, potentially increasing even further their advantage in the market, sometimes at the expense of startups and smaller companies (Rotman, 2023).

Moreover, foundation models are often trained with the help of "ghost workers", who provide human feedback to optimize reinforcement learning in order to prevent discriminatory or offensive responses being generated for end users. These workers are often employed on relatively low incomes in global south countries, which can serve to reinforce the inequitable global landscape of the AI industry (Perrigo 2023, https://ghostwork.info/). This adds up to a highly unequal business model for AI, where a small number of countries and firms develop and control a large share of these technologies, and are the ones that have the skills, infrastructure, investment, and data to advance innovation.

On the other hand, while bearing in mind that close to half of the global population does not have access to a fixed broadband service or are not able to use it effectively (ITU, 2022), for those with good connectivity, the democratizing capacity of access to knowledge and digital services that ChatGPT brings to millions of low-income people stands out. For example, its ability to facilitate autonomous learning is highlighted, or the possibility of reducing barriers of access to a research assistant between the global north and south (Baidoo-Anu & Owusu Ansah, 2023; Dowling & Lucey, 2023; Firat, 2023).

In the Recommendation, Policy Area 10 on Economy and Labour stresses the need to invest in reskilling and upskilling workers, providing them with the tools and education needed to integrate AI effectively. Placing emphasis on the ethical aspect of job transformation is equally as important as the technical aspects, or in the Recommendation's words: "Skills such as 'learning how to learn', communication, critical thinking, teamwork, empathy, and the ability to transfer one's knowledge across domains, should be taught alongside specialist, technical skills, as well as low-skilled tasks". In addition, competitive markets and consumer protection should be ensured in order to prevent abuse of dominant market positions.

Transparency, explainability and verifiability

A related concern is the fact that models such as ChatGPT are opaque both in relation to the data set that has been used to train them (OpenAI refused to disclose what data had been used to train GPT-4; Barr, 2023), and the workings of the system itself in how it derives its answers. As Paragraphs 39 and 40 of the Recommendation state: "Transparency aims at providing appropriate information to the respective addressees to enable their understanding and foster trust. Specific to the AI system, transparency can enable people to understand how each stage of an AI system is put in place, appropriate to the context and sensitivity of the AI system. It may also include insight into factors that affect a specific prediction or decision, and whether or not appropriate assurances (such as safety or fairness measures) are in place... [E]xplainability is closely related to transparency, as outcomes and sub-processes leading to outcomes should aim to be understandable and traceable, appropriate to the context."

With regard to the reliability of information provided using these tools, the outputs produced are often not accurate or up to date. There has been little disclaimer to this effect, particularly with the initial version of ChatGPT, which had a cut-off training date in 2021. By default, these tools are not able to verify the accuracy of the information provided. Furthermore, when prompted to provide references or citations, they often fabricate made-up resources to support their outputs.

Lack of transparency and verifiability could contribute to the spread of disinformation and misinformation. ChatGPT presents its outputs in a persuasive and authoritative manner and can thus trigger adverse outcomes by generating fictitious information (Hacker News, 2023; Qadir, 2022) and by facilitating the creation and distribution of disinformation campaigns, especially if combined with other generative AI to create deepfakes (Edwards, 2023).

The first problem results from the system seeking to create human-like text which is not based on curated sources of knowledge but on a statistical model that seeks to optimize the prediction of the next word in a sentence. Its model can therefore fulfil its objective without necessarily being truthful, in acts often termed "hallucinations". The second problem can lead to cybersecurity breaches when it is used in phishing campaigns (Business Standard, 2022) and could trigger serious challenges to democratic and social processes if used to polarize or mislead the public through propaganda

or misinformation campaigns (McGuffie & Newhouse, 2020).

In the UNESCO Recommendation on the Ethics of AI, Policy Area 9 on Communications and Information calls for Member States to improve access to information and knowledge, to respect and promote freedom of expression and diversity of viewpoints, and to promote digital literacy skills. Thus, paragraph 114 of the Recommendation calls on Member States to "*... invest and promote digital and media and information literacy skills to strengthen critical thinking and competencies needed to understand the use and implication of AI systems, in order to mitigate and counter disinformation, misinformation and hate speech*."

As previously mentioned, current LLMs have been shown to generate fictional information and made-up academic references in support of their claims. Providing transparency and explainability could involve, at the least, providing a list of real references for factual claims made in a response so that users can understand where the answers they are getting come from, and are better empowered to judge their level of truth, bias, and trustworthiness – while also, where relevant, giving credit to the creators of the content from which the tool is deriving its outputs.

Responsibility and accountability

HLEG experts consistently expressed dissatisfaction in relation to OpenAI's terms of use for ChatGPT, as responsibility for the output of the tool is delegated entirely to users users (OpenAI, 2023a). In other words, the terms of use are designed in a manner that shields the companies behind the tools from any responsibility for the accuracy and reliability of their outputs. Accountability is thus very hard to achieve, especially in the absence of governance frameworks that spell out exact accountability requirements. This delegation of responsibility is blind to the structural effects of ChatGPT, and obscures the responsibility of the humans managing OpenAI and the companies developing generative AI products in general.

Paragraphs 42 and 43 of the Recommendation state that "*the ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system*" and demand appropriate oversight and measures to ensure accountability for AI systems and their impacts.

In addition, the principle of Awareness and Literacy stresses the need for publicly accessible and human-readable information both about the terms of reference and general understanding of AI technologies, as well as about the value of data, including via ethics training.

A related question is whether these tools can be (legally or morally) considered authors or co-authors. The US Copyright Office has rejected one attempt to copyright AI artwork (The Verge, 2022), and *Nature* and *Science* recently confirmed that ChatGPT does not meet their standard to be considered an author (Stokel-Walker, 2023; Thorp, 2023). Despite this, several authors of journal articles have recently tried to claim ChatGPT as a co-author. This triggers questions about the moral and legal standing of AI as well as about responsibility and accountability in the case of, for example, errors made in AI-authored papers. The Editor-in-Chief of Science is clear on their position "*For the Science journals, the word "original" is enough to signal that text written by ChatGPT is not acceptable: It is, after all, plagiarized from ChatGPT.*" In the case of graphics generated by AI models, many artists have criticized the way that such models have been trained on their own works without consent, and Stability AI (the company that makes the AI tool Stable Diffusion) is currently being sued by Getty Images for copyright infringement.

Some argue that humans could be deemed as creators of inventions produced by AI with adequate human supervision (Iaia, 2022). Since AI cannot be considered creative due to the human footprint of copyright law, works generated by AI are typically considered the property of the human creator. However, the definition of "adequate" remains unclear.

On this point, Paragraph 68 of the Recommendation states unequivocally that "*AI systems should not be given legal personality themselves*", ruling out possibilities of this type.

Safety and security

Paragraph 27 of the Recommendation states that safety and security risks must be prevented and eliminated from the entire life cycle of AI systems. However, some uses of foundation models and ChatGPT, in particular, can cause significant damage that has not been foreseen or sufficiently addressed. The first security issue occurs when this tool is used to facilitate the distribution of content that may, by its very nature, be dangerous. For example, through a particular prompt, ChatGPT has been able to deliver instructions for building a dirty bomb (outrider.org, 2023). The HLEG also identifies a similar case in the use of the tool's programming capabilities to facilitate the creation of computer viruses, including malware, ransomware, spyware and the previously

mentioned phishing campaigns – potentially contributing to a range of cybersecurity threats and putting the tools to create them in the hands of a far greater number of people. Concerns were expressed about the short-term approach of the companies behind these technologies of creating patch solutions instead of addressing these problems in the architecture and design of the systems themselves.

Privacy and data protection

Paragraphs 32, 33 and 34 of the Recommendation address the issues of privacy and data protection as crucial elements in defending human dignity, autonomy, and agency. Significant importance is given to enforcing national and international law in the collection, use, sharing, storage and deletion of data and the adoption of adequate data protection frameworks and governance mechanisms. While this point has until recently been less present in public discussion, it is not clear that the practices of generative AI companies are geared towards protecting people's private information. With the right prompt, these systems could reveal data from their training data set, including providing personal information about individuals collected from the open internet that may never have been intended to be processed and made available in this way and within this use context.

Moreover, in March 2023, ChatGPT was briefly taken offline after experiencing a bug that allowed some users to see the titles from another user's chat history and may also have made visible the payment information of some subscribers (OpenAI, 2023b). Later in the same month, the Italian data protection authority blocked the use of ChatGPT, citing privacy concerns about the way it was gathering data as well as its lack of age verification, and opened an investigation into whether the tool is compliant with the General Data Protection Regulation. This block was lifted recently, but other European regulators have stated that they have similar concerns and are actively coordinating with the Italian authority on this matter (Mukherjee, Pollina & More, 2023). The investigation is ongoing at time of writing.

Human and environmental flourishing

Policy Area 5 on the environment and ecosystems urges Member States and companies to take responsibility for the direct and indirect environmental impacts of AI. Foundation models depend on enormous data processing, consuming huge amounts of energy. This goes against the goal of net zero emissions and actually contributes to rather than addresses global warming. Having said that, a comprehensive evaluation of the environmental impact of generative AI is yet to be provided, and the possible environmental benefits that particular applications of these models can have, have not yet been assessed.

In relation to human flourishing, at least two concerns have been raised by experts. These are: the degradation of social interactions; and the risk of affecting, in the long run, the cognitive abilities associated with literacy, including writing, understanding and critical thinking. Text generation tools raise the prospect of degrading social interactions and relationships if used widely to mediate human-human communication, for example, by automatically answering emails or responding to instant messages. This point has been identified as one of the pitfalls of using ChatGPT in education (Baidoo-Anu & Owusu Ansah, 2023), and some AI experts have made an explicit call to protect the human teacher-student relationship. Regarding the automation of writing, its role in the development of critical thinking and creativity has been highlighted, with the danger that writing could cease to be a creative and reflective space if the practice of editing text created by AI were normalized without safeguards in place (see, e.g., Puschak, 2023).

Education and research

In education and research, the debate has focused on the possibility of using ChatGPT to cheat on evaluations; on authorship and referencing scientific research; on the degradation of social relations in the educational process; and on the long-term effects on literacy-related competencies (Baidoo-Anu & Owusu Ansah, 2023; Cotton et al., 2023; Firat, 2023; Kasneci et al., 2023; Susnjak, 2022; Zhai, 2022).

With respect to scientific research, it is well-recognized that not only the web, but also non-peer-reviewed journals, contain inaccurate or outdated scientific theories and data. As outputs will be pieced together from potentially unreliable online sources, generative AI tools are open to presenting erroneous data and theories as established and accepted knowledge, in addition to the problem of hallucinated references explained above. Moreover, the ease with which LLMs can be used to generate sections of scientific papers could greatly increase the number of low-quality research papers in circulation and erode the quality and originality of scientific publications, which in turn could end up becoming part of the training data set for future LLMs.

Concerning education, the list of potential benefits for

students includes: facilitating personalized tutoring; providing automated essay grading and suggestions for improvement; allowing for rapid language translation; supporting autonomous learning, which could e.g. help students with disabilities; and enabling interactive, adaptive, asynchronous and remote learning experiences. On the negative side, it is highlighted that ChatGPT could reduce human interaction; that students reach only a limited understanding of the contents; that it reproduces social biases; that individuals could become dependent on the tool; that it could encourage dishonesty in evaluations; and could lead to the spread, particularly in low- and middle-income countries, of poor-quality, auto-generated, and uncurated learning materials and curricula that could contain factual errors, bias, and may not cover important or controversial areas of educational content. ChatGPT is further known to trigger cognitive biases in students. For example, it over-reproduces its "favourite" number, seven, in its answers, and over-justifies its incorrect responses (Azaria, 2023).

Some policy responses have emerged in response to these problems, including banning the use of ChatGPT completely or in evaluations, using other software tools to detect AI-generated text (roberta-base-openai-detector, 2022; Writer, 2023; although there are doubts about their effectiveness: Williams, 2023; Tate, 2023), returning to invigilated and oral exams, adjusting assignments and guidelines to integrate the use of ChatGPT, or even experimenting with integrating ChatGPT directly into pedagogical practice, for example by creating interactive assessments and games. Many of these responses require training for students and teachers and the setting of new guidelines, entailing the commitment of additional resources and potentially subtracting from time and resources available for teaching.

OpenAI has stated that they plan to develop a cryptographic watermark to identify ChatGPT's outputs as AI-generated, although this has not yet been implemented.

Conclusion

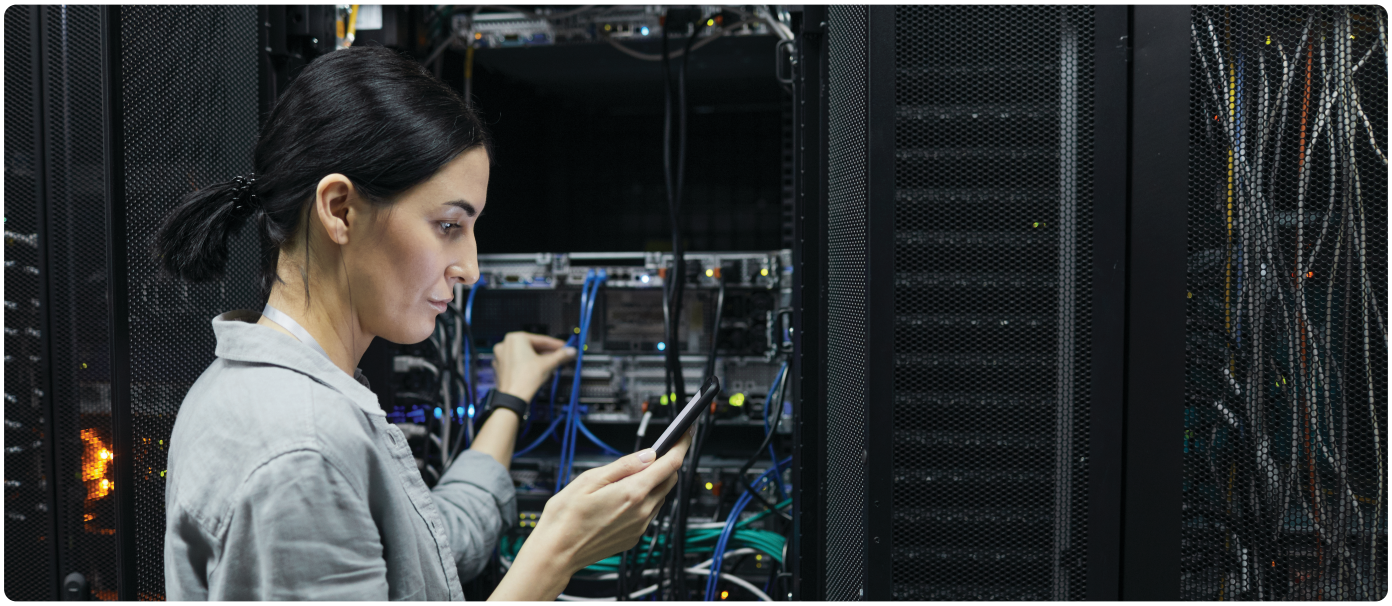
In these sections we have enumerated some of the concerns related to foundation AI models through the lens of the UNESCO Recommendation. As this analysis demonstrates, tools such as ChatGPT are not currently being designed, developed, and deployed in a manner that is compliant with the Recommendation.

The unleashing on the public of the current generation of "experimental" AI tools such as ChatGPT provides a prime example of why it is imperative for Member States to implement the Recommendation in order to identify, clarify, and mitigate the risks of harm from such models, and in so doing govern these models responsibly. This is particularly the case for subtle and longer-term harms that may be less apparent to the public or everyday users, such as environmental harms, bias, and impacts on critical thinking and creativity.

These concerns could have been identified during the design and development stages via a robust multistakeholder engagement process and ethical impact assessment – both key elements of the Recommendation – that could have helped pinpoint the potential for harmful outputs, and facilitated testing and implementing sufficient mitigation measures. Likewise, improving public awareness and literacy, another of the Recommendation's principles, could have prepared the public for negative impacts that could not be mitigated through technical means.

The Recommendation is intended to provide an ex-ante assessment for governments, companies, and other organizations to design, develop, deploy, and procure AI systems ethically and in line with human rights and fundamental freedoms in order to prevent harms from occurring in the first place. However, it also has application to AI models after they are deployed and as they continue to be iteratively updated. To continue the example of ChatGPT, based on the analysis above, OpenAI could take several actions to mitigate some of the risks identified, such as disclosing in full the data set used to train GPT-4, and ensuring that ChatGPT provides references to support any factual claims it makes in its responses. It should also include clear information to the user that they should not take outputs from the platform at face value, but maintain a critical perspective in order to maximize the contributions the tool can make, while controlling the downsides.

This analysis demonstrates the urgent need to examine generative models and their applications through the lens of the Recommendation's values, principles, and policy areas, and to use tools such as ethical impact assessment to improve future iterations in order to limit risks and harms and identify and promote the social benefits of such systems.



References

Abid, A., Farooqi, M., & Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 298-306. https://doi.org/10.1145/3461702.3462624

Alba, D. 2022. OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails. Bloomberg.Com. https://www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results

Azaria, A. 2023. ChatGPT: More Human-Like Than Computer-Like, but Not Necessarily in a Good Way.

Baidoo-Anu, D., & Owusu Ansah, L. 2023. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4337484

Barr, K. 2023. "GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery. Gizmodo. https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989

Benjamin, R. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. Polity.

Bhadani, R. 2022. Inherent Human-bias in Chat-GPT. MLearning.Ai. https://medium.com/mlearning-ai/inherent-human-bias-in-chat-gpt-ed803d4038fe

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. 2020. Language Models are Few-Shot Learners (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Business Standard. 2022. Hackers can use AI chatbot ChatGPT to write phishing emails, codes: Experts. https://www.business-standard.com/article/current-affairs/hackers-can-use-ai-chatbot-chatgpt-to-write-phishing-emails-codes-experts-122122000611_1.html

Cain, S. 2023. 'This song sucks': Nick Cave responds to ChatGPT song written in style of Nick Cave. The Guardian. https://www.theguardian.com/music/2023/jan/17/this-song-sucks-nick-cave-responds-to-chatgpt-song-written-in-style-of-nick-cave

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. B. 2023. ChatGPT Goes to Law School. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4335905

Cotton, D., Cotton, P., & Shipway, J. R. 2023. Chatting and Cheating. Ensuring academic integrity in the era of ChatGPT [Preprint]. EdArXiv. https://doi.org/10.35542/osf.io/mrz8h

Dowling, M., & Lucey, B. 2023. ChatGPT for (Finance) research: The Bananarama Conjecture. Finance Research Letters, 103662. https://doi.org/10.1016/j.frl.2023.103662

Edwards, B. 2023. Microsoft's new AI can simulate anyone's voice with 3 seconds of audio. Ars Technica. https://arstechnica.com/information-technology/2023/01/microsofts-new-ai-can-simulate-anyones-voice-with-3-seconds-of-audio/

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. (arXiv:2303.10130v4). arXiv. https://arxiv.org/pdf/2303.10130.pdf

Enoch, E. 2022. Will Chat GPT Replace Your Job As a Programmer? CodeX. https://medium.com/codex/will-chat-gpt-replace-your-job-as-a-programmer-3492ad2cf449

Firat, M. 2023. How Chat GPT Can Transform Autodidactic Experiences and Open Education? OSF Preprints. https://doi.org/10.31219/osf.io/9ge8m

Gozalo-Brizuela, R., & Merchan-Garrido, E. C. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models (arXiv:2301.04655). arXiv. http://arxiv.org/abs/2301.04655

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection (arXiv:2301.07597). arXiv. http://arxiv.org/abs/2301.07597

Hacker News. 2023. ChatGPT produces made-up nonexistent references | Hacker News. https://news.ycombinator.com/item?id=33841672

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/

Iaia, V. 2022. To Be, or Not to Be … Original Under Copyright Law, That Is (One of) the Main Questions Concerning AI-Produced Works. GRUR International, 71(9), 793-812. https://doi.org/10.1093/grurint/ikac087

Imgur. 2022. ChatGPT, an AI, interactively helps design a D&D adventure. Imgur. https://imgur.com/a/9cKhFO4

ITU. 2022. ITU DataHub. https://datahub.itu.int/data/?e=FRA&c=&i=19303

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. 2023. Is ChatGPT A Good Translator? A Preliminary Study (arXiv:2301.08745). arXiv. http://arxiv.org/abs/2301.08745

Johnson, K. 2023. The Efforts to Make Text-Based AI Less Racist and Terrible. Wired. https://www.wired.com/story/efforts-make-text-ai-less-racist-terrible/

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education [Preprint]. EdArXiv. https://doi.org/10.35542/osf.io/5er8f

Li, L., & Bamman, D. 2021. Gender and Representation Bias in GPT-3 Generated Stories. Proceedings of the Third Workshop on Narrative Understanding, 48-55. https://doi.org/10.18653/v1/2021.nuse-1.5

McGuffie, K., & Newhouse, A. 2020. The Radicalization Risks of GPT-3 and Advanced Neural Language Models (arXiv:2009.06807). arXiv. http://arxiv.org/abs/2009.06807

McQuillan, D. 2023. We come to bury ChatGPT, not to praise it. https://www.danmcquillan.org/chatgpt.html

Noble, S. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

O'Neil, C. 2016. Weapons of Math Destruction. Crown Books.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. OpenAI. https://openai.com/blog/chatgpt/

outrider.org. 2023. Could a Chatbot Teach You How to Build a Dirty Bomb? Outrider. https://outrider.org/nuclear-weapons/articles/could-chatbot-teach-you-how-build-dirty-bomb

Perrigo, B. 2023. Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. Time. https://time.com/6247678/openai-chatgpt-kenya-workers/

Powell, O. 2023. OpenAI confirms ChatGPT data breach. Cyber Security Hub. https://www.cshub.com/data/news/openai-confirms-chatgpt-data-breach

Qadir, J. 2022. Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. TechRxiv. https://doi.org/10.36227/techrxiv.21789434.v1

roberta-base-openai-detector. 2022. Roberta-base-openai-detector · Hugging Face. https://huggingface.co/roberta-base-openai-detector

Rotman, D. 2023. ChatGPT is about to revolutionize the economy. We need to decide what that looks like. MIT Technology Review. https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/

Samuel, S. 2022. A new AI draws delightful and not-so-delightful images. Vox. https://www.vox.com/future-perfect/23023538/ai-dalle-2-openai-bias-gpt-3-incentives

Seghier, M. 2023. ChatGPT: not all languages are equal. Correspondence. Nature, 615, 216.

Stokel-Walker, C. 2023. ChatGPT listed as author on research papers: Many scientists disapprove. Nature, 613(7945), 620-621. https://doi.org/10.1038/d41586-023-00107-z

Susnjak, T. 2022. ChatGPT: The End of Online Exam Integrity? (arXiv:2212.09292). arXiv. http://arxiv.org/abs/2212.09292

The Verge. 2022. The US Copyright Office says an AI can't copyright its art. https://www.theverge.com/2022/2/21/22944335/us-copyright-office-reject-ai-generated-art-recent-entrance-to-paradise

Thorp, H. H. 2023. ChatGPT is fun, but not an author. Science, 379(6630), 313-313. https://doi.org/10.1126/science.adg7879

Walker, J. 2022. ChatGPT is multilingual but monocultural, and it's learning your values. https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/

Writer. 2023. AI content detector. Writer. https://writer.com/ai-content-detector/

Zhai, X. 2022. ChatGPT User Experience: Implications for Education. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4312418

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. 2023. Exploring AI Ethics of ChatGPT: A Diagnostic Analysis (arXiv:2301.12867). arXiv. http://arxiv.org/abs/2301.12867

unesco
United Nations
Educational, Scientific
and Cultural Organization

Social and Human Sciences Sector
7 Place de Fontenoy
75007 Paris, France

ai-ethics@unesco.org

on.unesco.org/EthicsAI

Follow us
@UNESCO #AI #AIEthics

Read UNESCO's Recommendation on the Ethics of Artificial Intelligence

SHS/2023/PI/H/12
https://doi.org/10.54678/BGIV6160