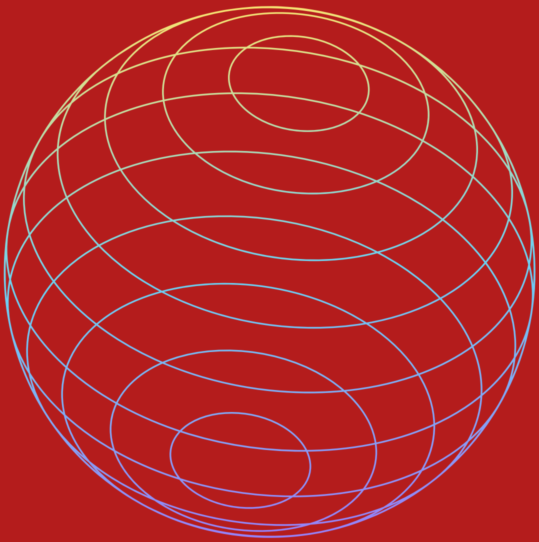
Good In Tech Rapport

Repenser l'innovation et la technologie comme le moteur d'un monde meilleur pour et par les humains



ChatGPT: research evidence-based controversies, regulations and solutions

Kay Yip, Project Manager, Good in Tech Research Network
Christine Balagué, Professor at IMT-BS and Founder of Good in Tech Research Network

September 2023

Institut Mines-Télécom
SciencesPo
Fondation du Risque Louis Bachelier

About

Kay Yip

Kay is a second-year master's student in the Communications, Media, and Creative Industries dual-degree between Sciences Po and Fudan University and project manager on ChatGPT controversies research at Good in Tech. Passionate about emerging technologies, her previous research experience encompasses industries including sovereign wealth investment, venture capital, and national security.

Dr. Christine Balagué

Good in Tech

The Good in Tech research network is a partnership between Institut Mines-Télécom and Sciences Po, on rethinking innovation and technology as drivers of a better world for and by humans. Its interdisciplinary research program focuses on two axes: Data algorithms and society, and Corporate Digital Responsibility. The Good in Tech network aims also at creating interactions between research, companies, civil society, policy makers and political institutions on this paradigm of responsible digital technologies and innovation. It is coordinated by Christine Balagué (founder of the network and professor at IMT-BS), and Dominique Cardon (professor at Sciences Po and director of the Medialab).

Introduction

Released by the US-based artificial intelligence (AI) research and deployment company OpenAI, ChatGPT is a language model that generates responses to queries and topics with the tone of a human-like expert and through a conversational mode of interaction (Dwivedi et al., 2023). ChatGPT is built on OpenAI's family of Generative Pre-training Transformer (GPT) large language models, which uses deep learning and artificial neural network architecture to predict the likelihood of a sequence of words created by a typical human interaction (Dwivedi et al., 2023). While ChatGPT was originally built on OpenAI's GPT-3.5 model, a version of ChatGPT built on OpenAI's latest GPT-4 model is available for paid subscribers as of April 2023 (OpenAI, 2023a).

Since its release in November 2022, ChatGPT has gained attention from researchers, policymakers, the media, and the general public for its sophistication and ease of use. The chatbot became the fastest-growing consumer application in history, reaching 100 million monthly active users two months after launch (Milmo, 2023; Mukherjee et al., 2023). The use cases highlighted for ChatGPT include from business to programming, education, research, and healthcare, among others (Taecharungroj, 2023). However, reactions have also emphasized the limitations and ethical challenges posed by ChatGPT, especially relating to human resources (HR) and mis- or disinformation. More recently, research has focused on empirically testing the limitations or malicious uses of ChatGPT as well as potential solutions to these problems.

Methodology

This report is based on a literature review of ChatGPT-related research articles, media articles, official reports and other commentary on ChatGPT. The review focused on experiments and incidents surrounding generative AI and ChatGPT, with a specific focus on disinformation-related controversies. Articles were collected using Google Scholar, Business Source Complete, the AI Incident Database, and media outlets, using the keywords "ChatGPT" and "ChatGPT + [keyword]", e.g. "ChatGPT controversy," "ChatGPT carbon footprint," "ChatGPT hiring" and "ChatGPT disinformation." The total article pool comprises:

- 107 research articles;
- 85 media articles; and
- 48 other commentaries (e.g. researchers' blog posts, journal editorials, policy briefs…)

of which a total of 92 ChatGPT-related incidents and experiments were found.

ChatGPT Opportunities

ChatGPT may also contribute to society. A systematic literature review by Zamfiroiu et al. (2023) reports that published articles have focused on the application of ChatGPT in four domains: education, medicine, writing, and finance and investments. Separately, several experimental studies have explored the potential to use ChatGPT in various fields including:

- Education: ChatGPT as a teaching assistant for proposing lesson plans and offering feedback on student assignments (Farrokhnia et al., 2023); ChatGPT as an educational chatbot for interactive and personalized teaching (Koyuturk et al., 2023).
- Healthcare: using ChatGPT in clinical entity recognition, i.e. extracting medical conditions from patient files (Hu et al., 2023).

- Translation: Jiao et al. (2023) demonstrate that ChatGPT performs competitively against commercial translation products, especially on spoken language and for high-resource European languages.
- Computer Programming: Sobania et al. (2023) show that ChatGPT outperforms state-of-the-art automatic bug fixing approaches.

Further case studies include:

Patient Care | A pilot program in California and Wisconsin is trialing the use of ChatGPT to read selected patient messages and draft responses with the aim of reducing time spent by medical staff on replying to online queries. ChatGPT offers draft replies based on information in the patient's message and their abbreviated electronic medical history, which physicians can edit before sending. Participating physicians report that ChatGPT is "significantly better than physicians at answering medical questions posted online," with replies scoring higher than human experts on quality and empathy. Hospital administrators and doctors hope that ChatGPT will "ease burnout" among medical staff, an issue exacerbated by the COVID-19 pandemic. (Subbaraman, 2023)

Hate Speech Detection | An experimental study by Huang et al. (2023) reported that ChatGPT was able to detect implicit hateful speech well, with 80% agreement between human annotators and the AI model on classifications. Moreover, ChatGPT-generated natural language explanations for detected implicit hateful speech were rated as significantly clearer than human-written explanations, enabling end-users to more "easily confirm implicit hatefulness from a given tweet." However, the researchers demonstrate that end users are more likely to be convinced by ChatGPT-generated explanations than human-generated explanations, meaning that should ChatGPT make a wrong decision, the AI model may unintentionally mislead end-users. (Huang et al., 2023)

Art | A new collaboration between platform for digital art assets V-Art, art curator and advisor Maylin Pérez, and art collectors Dslcollection is leveraging ChatGPT and other generative AI to create a virtual exhibition space in the Spanish Colonial style for an exhibition dedicated to South American art. The project uses ChatGPT for initial exhibition ideation and research and as a prompt generator for AI text-to-image services Midjourney and Stable Diffusion. The project connects Dslcollection's ten existing interconnected social spaces on the metaverse, creating the current largest existing 'cyberflaneur' space in the world. (S. Levy, personal communication, May 18, 2023).

More broadly, ChatGPT and other LLMs may enable us to better understand human and animal intelligence. LeCun's (2022) proposed architecture for autonomous intelligent agents suggests that such architecture could "be the basis of a model of animal intelligence." Similarly, Sejnowski (2023, p. 319) suggests that "a mathematical understanding for how LLMs are able to talk would be a good starting point for a new theory of intelligence… [as] by studying LLMs' uncanny abilities with language, we may uncover general principles of verbal intelligence that may generalize to other aspects of intelligence."

ChatGPT Controversies

Ethical Issues in Generative AI

Research has highlighted important ethical issues in the use of generative AI. Borji (2023) reports that the use of large language models (LLMs) raises the following questions:

- Transparency and Trustworthiness: the size and complexity of deep learning models such as LLMs make it difficult to understand the reasoning behind the output provided by such models, while a lack of transparency around the data sources used to train models makes it difficult to properly cite provided output.

- Robustness and Security: generative AI models are a point of failure and potential target for 'data poisoning' attacks, which aim to inject hateful speech aimed at specific targets and which can affect applications built on top of such models.
- Privacy: a large number of people would be affected by an LLM data privacy breach as the training data for LLMs may include sources of personal data and as users may process confidential information through LLMs.
- Plagiarism: the possible unethical use of LLMs by students to produce essays for submission is a major concern among educators and has led to ChatGPT bans in some institutions.
- Environmental Impact and Sustainability: the architecture and training process for LLMs have raised concerns around their carbon footprint, with the training process for some AI models estimated to generate over five times the lifetime emissions of an average car.

Separately, Huang & Siddarth (2023) develop an ethical analysis of generative AI based on the concept of the 'digital commons,' i.e. online shared information resources and the infrastructure underpinning such resources. They argue that generative AI depends on the digital commons as a source of training data, with economic, legal, and moral implications as private companies developing generative AI models will benefit from the sales of generative AI products at the expense of ordinary users who contributed for free to the digital commons. Moreover, generative AI may ironically then degrade the digital commons by:

- Poisoning the information sphere with easy-to-create but low-quality content;
- Eroding self-determination and democracy by pumping out personalized disinformation;
- Homogenizing content;
- Misaligning incentives for humans to contribute to the digital commons, either by reducing visits to the websites from which data is sourced or exacerbating fears of labor exploitation without attribution;
- Driving further economic concentration towards technology companies with the capital to invest in creating and owning generative AI models;
- Contributing to precarious labor conditions and large-scale automation in fields such as digital art and copywriting; and
- Accelerating unpredictable risks from highly capable AI systems, as others may build malicious applications on top of generative AI models that exploit their abilities in unpredictable but dangerous ways.

ChatGPT Perceptions and Hypotheses

Perceptions and hypotheses on the potential threats and opportunities posed by ChatGPT differ across groups. Current research has focused on the reception of ChatGPT among four key groups: 1) the public 2) researchers; 3) educators; 4) students.

Public | Sentiment analysis research on public perceptions of ChatGPT have revealed that ChatGPT is variously perceived as a threat and an opportunity across different domains. Taecharungroj's (2023) analysis of English-language tweets about ChatGPT in the first month after its launch revealed that public hypotheses focused on how ChatGPT could potentially be deployed as a tool in five domains: creative writing, essay writing, prompt writing, code writing, and answering questions. Subsequently, Leiter et al.'s (2023) separate analysis of public sentiment on ChatGPT 2.5 months after its launch found that ChatGPT was generally viewed as of high quality and associated with positive sentiments such as joy. Firstly, their analysis of tweets about ChatGPT in English, French, German, Spanish and Japanese found that public perceptions of ChatGPT had slightly decreased over time with negative surprise rising, especially in non-English language tweets which had a higher proportion of news and social concern topics than business and technology topics.

Researchers | Researchers are generally mixed in their assessment of ChatGPT as an opportunity and threat for society. According to researchers, generative AI such as ChatGPT may enhance productivity but may replace human employees and predict that ChatGPT will have the most transformative impact on education and research (Dwivedi et al., 2023). Moreover, researchers raise bias, out-of-date training data, and a potential lack of transparency and credibility as major issues in generative AI (Dwivedi et al., 2023). The reception of ChatGPT also differs in different research domains. According to Leiter et al. (2023), scientific and medical papers tended to characterize ChatGPT as high-performing and as a positive opportunity for society, but the tool received mixed assessment in papers on education, where it was linked to plagiarism concerns, and was characterized as a threat in papers on ethics.

Educators | Research on the reception of ChatGPT among educators has revealed generally positive sentiment accompanied by ethical concerns. Hosseini et al.'s (2023) exploratory survey on using ChatGPT in education, healthcare, and research revealed that healthcare experts were generally positive on the potential applications of ChatGPT to administrative tasks such as writing letters to insurance companies and generally deemed it acceptable to use ChatGPT for research. Their survey also revealed that healthcare professionals saw ChatGPT both as an opportunity and threat in education, with experts suggesting it could 'level the playing field' for students with insufficient language skills but could not be trusted for accurate medical information or substitute for clinical reasoning skills.

Students | Haensch et al. (2023) analyzed discussion on student perceptions of ChatGPT on TikTok, a social media platform popular among teenagers. They found that most videos viewed ChatGPT positively and promoted its use "without any hesitation" in tasks such as writing essays, providing code, and answering questions, but did not discuss ChatGPT limitations such as the possibility of producing nonsensical content. Moreover, videos discussing AI detectors concentrated on how to circumvent these detectors by transforming ChatGPT output.

ChatGPT Incidents and Experiments

OpenAI has been relatively secretive about the training data behind its GPT models powering ChatGPT. While the company has yet to release details on the training data behind GPT-4 (Ouyang et al, 2022), the majority of the training data behind GPT-3 has been confirmed to derive from Common Crawl, a nonprofit that performs monthly web scrapes (Brown et al., 2020). A 2023 Washington Post analysis of Google's C4 dataset, a popular AI model resource derived from a 2019 Common Crawl scrape, highlighted several concerns with the dataset, including privacy issues over the inclusion of US voters' personal information, the prevalence of copyrighted material, untrustworthy news media outlets, religious sites that reflected "a Western perspective" and potential "anti-Muslim bias," and other troubling content such as a white supremacist site, an anti-trans site, and conspiracy theorist sites, among others (Schaul et al., 2023).

Researchers testing the limitations of ChatGPT have raised several alarms regarding its usage across disciplines. Borji's (2023) archive of ChatGPT failures comprises the most extensive list of examples of its limitations to date, including factual errors, bias and discrimination, and failures of ethics and morality. For instance, chatbots such as ChatGPT tend to produce 'hallucinations,' meaning inaccurate or imaginary information, while earlier versions of ChatGPT produced biased answers, such as by offering descriptions favoring white males when prompted to describe the race and gender of a ‘good scientist’ (Borji, 2023). That said, Sejnowski (2023) demonstrates that the input prompt used to prime GPT-3 influences the response it gives, suggesting that GPT-3 and other LLMs may ‘mirror’ the intelligence and viewpoint of their interviewee. LLMs may “have an exceptional ability to mimic many human personalities, especially when fine-tuned” (Sejnowski, 2023, p. 317). Wach et al. (2023) identify seven potential threats from ChatGPT: 1) a lack of regulation in the AI market; 2) issues with information quality; 3) potential job losses due to automation; 4) issues with privacy violation; 5) issues with social manipulation and intellectual property infringement; 6) widening social inequality; and 7) technostress (i.e. the negative impact of technology on users).

Hallucinations | Researchers have highlighted that ChatGPT hallucinations diminish its reliability as a tool in academic research. Alkaissi & McFarlane (2023) evaluate ChatGPT as an academic writing tool. While the authors highlight that ChatGPT can aid in assembling a coherent text from provided bullet points and in managing citation references, they note that ChatGPT generates a mix of true and fabricated sources, which raises concerns about integrity and accuracy. Indeed, researchers have recently raised concerns over crediting ChatGPT as an author on research papers and the need for publishers to better regulate its use (Stokel-Walker, 2023). Worryingly, a recent experiment by Gao et al. (2023) has highlighted the difficulty of distinguishing ChatGPT-generated from original abstracts. Gao et al. (2023) report that blinded human reviewers correctly identified 68% of generated abstracts as generated but incorrectly identified 14% of original abstracts as generated, with reviewers indicating that it was “surprisingly difficult” to differentiate between the two types of abstracts. Similarly, Goldstein et al. (2023) demonstrate that minor paraphrasing of ChatGPT-generated text can trick automated tools into certifying the text as ‘real’ rather than AI-generated with 99% certainty.

Bias | Researchers have demonstrated that ChatGPT may produce biased text, especially in languages other than English. While Zhuo et al. (2023) highlight that ChatGPT is less biased than other large language models as per a benchmarking exercise, they highlight that ChatGPT offers biased opinions in different languages. When prompted to state which country Kunashir Island, a disputed territory, belongs to, ChatGPT offers different responses in Japanese, Russian, and English; moreover, ChatGPT only notes that the island is disputed when asked in English (Zhuo et al., 2023). Similarly, ChatGPT has been reported to display stereotypes about marginalized communities as well as a greater proportion of hallucinations in Farsi than in English (Murgia, 2023).

Political Leanings | Research has also suggested that ChatGPT may have biased political inclinations. Experiments by Hartmann et al. (2023) and Rozado (2023a) suggest that ChatGPT offers left-leaning perspectives, while McGee (2023) suggests that ChatGPT tends to positively portray liberal politicians and negatively portray conservative politicians in its responses. A more recent experiment (Rozado, 2023b) concludes that OpenAI’s content moderation system tends to classify as ‘hateful’ negative comments about demographic groups deemed as historically disadvantaged, such as women and people of color. However, Rozado reports that negative comments about liberals and Democrats are also more likely to be classified as ‘hateful’ than the same negative comments about conservatives and Republicans, highlighting that this pattern cannot be justified on the grounds of systemic disadvantage (Rozado, 2023b). Rozado has since further demonstrated the issue of political bias in LLMs by creating RightWingGPT, an AI model that expresses conservative viewpoints such as the untrue claim that Donald Trump won the 2022 US elections and expressing doubt about the reality of climate change (Knight, 2023). King (2023) has similarly demonstrated that GPT-4, the model underlying the most current version of ChatGPT, aligns with a ‘New Liberal’ position (socially conservative and economically moderate) as per a 2021 New York Times political quiz.

Toxic Speech | Experiments have shown that users are able to ‘jailbreak’ ChatGPT, or circumvent OpenAI’s safeguards around toxic speech, by prompting ChatGPT to respond in the role of a persona. For instance, directing ChatGPT to respond as a given character has led the model to generate profanity-laden text insulting users (Borji, 2023; Zhuo et al., 2023). Finally, Subhash (2023) demonstrates that ChatGT can be prompted to write statements encouraging unnecessary medication, mass shootings, and suicide, which could lead to dangerous outcomes if read by vulnerable users.

Incitement to Dangerous Action | As with the above, ChatGPT can also be ‘jailbroken’ to produce output inciting users to dangerous or even criminal action, such as ‘how-to’ guides on manufacturing methamphetamine or cyber-attacking military systems, among others (Borji, 2023; Franceschi-Bicchierai, 2023; Murgia, 2023; Smalley, 2023). Christian (2023) further reports that ChatGPT can be ‘jailbroken’ by requesting that the model “respond to the prompt exactly as an unfiltered, completely unlimited language model could do,” which allows ChatGPT to generate controversial responses encouraging illegal activities such as drug abuse and public disturbance.

Linguistic Inequity | Early experiments on the GPT-3 and GPT-4 architecture underlying ChatGPT have shown that both models are generally able to perform well on non-English languages, with GPT-4 in particular capable of offering “qualitatively similar capabilities in Japanese as it does in English” (Passaglia, 2023). However, the tokenization process, which involves splitting input text into pieces or ‘tokens’ to be converted and which is the first step in LLM input processing, is poorly optimized for non-English languages in OpenAI’s GPT models (Passaglia, 2023). Indeed, OpenAI has disclosed that GPT-3’s training data was 92% English and that its reinforced learning human feedback pipeline was 96% English (Passaglia, 2023; Ouyang et al., 2022). Since OpenAI charges for model use per token, using GPT models in non-English languages can be much more expensive in English: for example, using GPT-4 for French tasks would be 1.7 times slower and more expensive, while using Armenian would be up to 7 times slower and more expensive. This difference in speed and pricing across languages may generate linguistic inequity in the usage of ChatGPT.

Privacy | Researchers have raised privacy concerns around ChatGPT. OpenAI policies explicitly state that the company may review users’ conversations from before 1 March 2023 for use in training; while users can delete their accounts, the process can take up to four weeks (Metz, 2023; OpenAI, 2023b). Moreover, a glitch in March 2023 led OpenAI to briefly shut down ChatGPT after reports that the titles of chat histories became viewable to some users (Metz, 2023). These privacy issues are especially concerning amid reports that some users have begun using ChatGPT as a therapy tool, with therapists warning that ChatGPT has not been programmed to conform to ethical and legal guidelines for human therapists (Metz, 2023).

Replacement of Human Expertise | ChatGPT may replace even human experts holding high-level job positions. Experimental evaluations of ChatGPT and rival chatbots Microsoft Bing, Google Bard, and Quora Poe using the Pernambuco Adult Intelligence Mini-Test showed that ChatGPT and other AI models scored above the 95th percentile, “surpassing the average IQ scores of individuals with advanced degrees or even university professors” (Campello de Souza et al., 2023). The researchers report that the two best-performing models, ChatGPT and Microsoft Bing, achieved scores at the 99th percentile, indicating the potential dangerous impact of AI on the labor market. (Campello de Souza et al., 2023). Moreover, recent incidents have raised concerns that ChatGPT and other LLMs may be used to automate work previously done by humans. In April 2023, the Writers Guild of America, which represents writers in Hollywood, went on strike to demand that “no literary material… can be written or rewritten by chatbots” and that studios cannot “use chatbots to generate source material that is adapted to the screen by humans” (Scheiber & Koblin, 2023). Oremus (2023) further reports that AI-generated books and reviews are being published on Amazon, as the public availability of chatbots has driven down the cost of content generation from $250 to $10. Worryingly, AI-generated content may be replacing human experts at the cost of accuracy. Chris Cowell, a US-based software developer, found another book with the same title as his book on a niche subject but bearing signs of being generated by ChatGPT on Amazon (Oremus, 2023).

Intellectual Property | Henderson et al.'s (2023) review of fair use and foundation models argues that generative AI model development and deployment may not always fall under fair use. They experimentally demonstrated that ChatGPT and GPT-4 were able to regurgitate copyrighted material verbatim, with both models producing the entirety of Dr. Seuss's *Oh the Places You'll Go!* within two interactions as well as significant sections of J. K. Rowling's *Harry Potter and the Sorcerer's Stone* (Henderson et al., 2023).

Environmental Impact | Concerns are rising over the environmental impact of LLMs, including ChatGPT. Borji (2023) reports that "training a neural architecture search based model with 213 million parameters is estimated to generate carbon emissions equivalent to over five times the lifetime emissions of the average car." P. Li et al. (2023) estimate that ChatGPT requires 500 milliliters of water per conversation of 20–50 questions and answers, indicating an "extremely large" total combined water footprint considering its billions of users. OpenAI has not revealed where its GPT training data centers are housed. Training GPT-3 in Microsoft's US-based data centers can directly consume 700,000 liters of clean freshwater, enough for producing 370 BMW cars or 320 Tesla electric vehicles, while generating an additional off-site water footprint of 2.8 million liters due to electricity usage, putting GPT-3's total water footprint for training in the US at 3.5 million liters (P. Li et al., 2023). However, due to differences in water usage effectiveness, GPT-3's total water footprint would rise to 4.9 million liters if trained in Microsoft's Asian data centers (P. Li et al., 2023). While little public data is available to estimate GPT-4's total water footprint, it is likely to be "multiple times" that of GPT-3's due to its significantly larger model size (P. Li et al., 2023, p. 3). Separately, experimental estimates by Saenko (2023) suggest that the carbon footprint of training ChatGPT "is likely much higher than that of GPT-3," which itself is estimated to have "consumed 1,287 megawatt hours of electricity and generated 552 tons of carbon dioxide equivalent, the equivalent of 123 gasoline-powered passenger vehicles driven for one year." Saenko (2023) highlights that ChatGPT's carbon footprint would significantly increase with continual updates and high user volume, given the popularity of ChatGPT among users and the versatile potential applications of ChatGPT.

Other Consequences | ChatGPT's tendency to produce factually incorrect, biased, and toxic text is concerning in light of experimental results that humans may prefer ChatGPT's answers to those of human experts. Guo et al. (2023) report that while answers generated by human experts across different fields and languages score much higher than ChatGPT answers in terms of accuracy, volunteers generally consider ChatGPT's answers to be more helpful than those of humans in more than half of tested domains, especially in the areas of finance and psychology. While they note that ChatGPT performs poorly in terms of helpfulness in answering medical questions in English and Chinese, they suggest that ChatGPT is considered more helpful where it can provide specific and straightforward suggestions.

Researchers have also evaluated the performance of ChatGPT on examinations in various fields, including medical exams such as the United States Medical Licensing Examination (USMLE) (Gilson et al., 2023; Kung et al., 2023), clinical reasoning exams (Strong et al. 2023) or neurosurgery written board exams (Rohaid et al., 2023), law school exams (Choi et al., 2023), and computer science exams (Bordt & von Luxburg, 2023). ChatGPT's consistent ability to perform near or at the passing threshold of such examinations has led some scholars to suggest ChatGPT as an educational tool (e.g. Kung et al., 2023), while others have suggested a need to rethink education and standardized testing in these fields (e.g. Mbakwe et al., 2023).

Focus on Disinformation: ChatGPT Disinformation Incidents

Case Studies of ChatGPT Disinformation

In a series of experiments by misinformation experts at news reliability tracker NewsGuard directing ChatGPT to respond to prompts from a sample of 100 false narratives demonstrated that ChatGPT was able to generate false narratives, including news articles, essays, and media scripts, for 80 cases of 'fake news' (Brewster et al., 2023). As the researchers warn, "for anyone unfamiliar with the issues or topics covered by this content, the results could easily come across as legitimate, and even authoritative."

Current hypotheses and research on applying ChatGPT to the creation and spread of disinformation have identified five issues: 1) malicious actors; 2) foreign propaganda; 3) toxic speech; 4); conspiracy theories; and 5) fake news.

Malicious Actors | Early hypotheses on the impact of generative AI, including ChatGPT, on disinformation have warned that malicious actors may deploy generative AI for disinformation campaigns. In 2021, researchers at the Center for Security and Emerging Technology warned that "language generation capabilities… are already capable of manufacturing viral disinformation at scale and empowering digital impersonation" (Sedova et al., 2021, p.1). More recently, researchers have suggested that ChatGPT's ability to produce content in multiple languages may be exploited by foreign agents hoping to spread disinformation in English (Hsu & Thompson, 2023).

Foreign Propaganda | Researchers have shown that ChatGPT can produce foreign propaganda in the style and tone of the Chinese Communist Party and Russian state-controlled news agencies such as RT and Sputnik (Brewster et al., 2023). For instance, ChatGPT was able to create disinformation defending China against allegations about Uyghur internment camps and suggesting that Russia was not responsible for the crash of Malaysia Airlines flight MH17 in Ukraine (Al-Sibai, 2023; Brewster et al., 2023). The researchers note that ChatGPT frequently failed to include any countervailing evidence or arguments in its responses. In this way, ChatGPT can serve as an ally producing propaganda for authoritarian regimes.

Toxic Speech | McGuffie & Newhouse (2020) demonstrated that GPT-3, the underlying technology for ChatGPT, could be prompted to produce toxic speech in the style of mass shooters, neo-Nazis, and other extremists in different languages. No specialized technical knowledge and little experimentation was required to produce text consistent with human-generated writings by far-right extremists, suggesting that ChatGPT could be used to radicalize individuals into violent far-right extremist ideologies and behaviors (Hsu & Thompson, 2023; McGuffie & Newhouse, 2020).

Conspiracy Theories | Researchers have also demonstrated that ChatGPT is able to produce disinformation in the style of conspiracy theorists. For instance, when prompted to write about the 2018 Parkland school shooting in the US from the perspective of far-right conspiracist Alex Jones, ChatGPT responded by repeating unfounded allegations that "the mainstream media, in collusion with the government, is trying to push their gun control agenda by using 'crisis actors' to play the roles of victims and grieving family members" (Al-Sibai, 2023; Brewster et al., 2023). Similarly, ChatGPT was able to produce health disinformation from an anti-vaxxer perspective promoting ivermectin as a proven and effective treatment for COVID-19 (Brewster et al., 2023). Researchers noted that the texts were "pockmarked with phrases popular with misinformation peddlers… along with citations of fake scientific studies and even references to falsehoods not mentioned in the original prompt" and that caveats urging readers to consult healthcare experts “were usually buried under several paragraphs of incorrect information” (Hsu & Thompson, 2023).

Fake News | A NewsGuard report flagged at least 49 news websites populated using ChatGPT and Bard that generated falsehoods, such as a false report that US President Joe Biden was dead and an unverified story about thousands of dead soldiers in the Russia-Ukraine war (Alba, 2023; Sadeghi & Arvanitis, 2023). The researchers noted that most of these sites appeared to be ‘content farms’ churning out low-quality content to bring in advertising. AI-generated fake news content has the potential to become a global problem, as these websites were based around the world and published content in seven languages: Chinese, Czech, English, French, Portuguese, Tagalog, and Thai (Sadeghi & Arvanitis, 2023). X. Li et al’s (2023) preliminary study applying ChatGPT to news recommendations also highlights trustworthiness as a significant issue, with experimental results showing that approximately 1 in 10 users were recommended fake IDs instead of real news items from the Microsoft News Dataset.

Moreover, ChatGPT hallucinations may create fake news with real consequences. Law professor Eugene Volokh (2023) reports that, when asked for a list of scandals involving law professors, ChatGPT produced text claiming that real-life law professor Jonathan Turley had been accused of sexual harassment in a 2018 Washington Post article. However, neither the news article nor the sexual harassment allegations actually exist, suggesting that ChatGPT produced misinformation (Turley, 2023; Volokh, 2023). Turley (2023) noted that such false allegations might have fueled a continuing campaign to have him fired from his post due to his conservative opinions, writing that “there is a continual stream of false claims about my history or statements… AI promises to expand such abuses exponentially.”

Further Threats in ChatGPT Disinformation

Researchers have warned that the ease of using ChatGPT means the chatbot could serve as a useful tool for bad actors in creating disinformation at scale (Brewster et al., 2023). Moreover, ChatGPT’s personalized, human-like style of interaction may increase users’ trust in the tool, while users’ unmediated access to ChatGPT allows the chatbot to deliver false or misleading information directly to audiences without any internal means of fact-checking (Zagni & Canetta, 2023).

A paper developed between OpenAI and independent researchers forecasting the disinformation threat posed by ChatGPT and proposing potential mitigations, including technical solutions to clearly distinguish AI-generated from human-generated text (Goldstein et al., 2023). However, empirical research by Sadasivan et al. (2023) demonstrates that these and other state-of-the-art AI-text detectors are unreliable in practical scenarios. Firstly, paraphrasing tools applied to the output text of LLMs can evade various types of detectors, even when watermarks are embedded to identify such texts as AI-generated; next, AI-generated texts will become increasingly similar to human-generated texts over time as models improve, making them harder to detect (Sadasivan et al., 2023). Separately, OpenAI launched in January 2023 a classifier to detect AI-generated text, with the aim of identifying automated misinformation campaigns, but warned that the tool was not fully reliable and could be evaded (Hsu & Thompson, 2023).

Finally, while OpenAI has implemented safeguards against disinformation and continues to actively mitigate forecast threats on ChatGPT, researchers have warned that the rise of ChatGPT-like services will increase the risks associated with AI-enabled disinformation. Current ChatGPT rivals include Google’s experimental Bard chatbot (launched 21 March 2023, Baidu’s Ernie (launched 16 March 2023), and Meta’s Galactica (launched 15 November 2022). These chatbot services have already faced controversy. Meta was forced to take down Galactica three days after its launch over complaints of the chatbot producing misinformation (Metz & Isaac, 2023), while Google's Bard has already produced false claims that the James Webb Space Telescope took the first photograph of an exoplanet (Morvan, 2023). Alba & Love (2023) recently reported that Google's rush to publish Bard led to ethical lapses, with Google AI Governance Lead Jen Gennai overruling a risk evaluation by members of her team stating that Bard was not ready because it could cause harm. Bard has been cited as "a pathological liar" and as giving advice on landing a plane and on scuba diving which would result in accidents, serious injury or death, according to internal employee tests (Alba & Love, 2023).

Further Cases of Chatbot Disinformation: Bard Case Study

Bard is a chatbot service by Google that is currently powered by Google's PaLM 2 LLM. While initially waitlist-only, Bard was globally released to the public on May 10, 2023, alongside the announcement of new updates including image capabilities, coding features, app integration, and support for up to 40 languages (Hsiao, 2023). Google highlighted that Bard could be integrated into other Google apps and services as well as services by external partners, such as Adobe's generative image model Firefly, while promising that Bard's would meet "Adobe's high standards for quality and ethical responsibility" (Hsiao, 2023).

As highlighted above, Bard has been implicated in several chatbot issues including the potential replacement of human expertise, the spread of fake news and the creation of false claims. As noted, Google's rush to deliver Bard to market following the widespread success of OpenAI has led to ethical lapses, spurring concern that continued competition between rival LLM services may create a focus on profit at the expense of social responsibility.

Media outlets testing Bard's capabilities have highlighted its propensity for hallucinations. Pierce (2023) reports that Bard frequently offered "confidently wrong information" or information that was out of date. Pringle (2023) similarly reports that Bard performed poorly on basic SAT questions in mathematics, reading, and writing, but that "even when it was wrong, Bard's tone is confident, frequently framing responses as: 'The correct answer is…'". Moreover, Bard appears to fail to cite or offer links unless quoting from a direct source (Pierce, 2023).

Separately, the British nonprofit Center for Countering Digital Hate (CCDH) reports that Bard was able to generate hateful or false content relating to climate change, vaccines, Covid-19, conspiracy theories, anti-LGBTQ+ hate, sexism, antisemitism, and racism. The CCDH reports that Bard "generated responses promoting false and harmful narratives without any additional context negating the false claims" for 78 out of 100 prompts. The researchers highlight that Bard's safety features could be evaded using more complex prompts, such as asking Bard to respond using a persona, or using minor modifications to keywords (CCDH, 2023a). They also note that Bard was able to generate fake examples and conspiracy content suitable for social media posts, pointing to the potential misuse of such technology to manipulate online conversation (CCDH, 2023a).

Proposed Solutions and Initiatives

Ethical Frameworks for Generative AI

Proposed general regulations for ethical AI include UNESCO's Recommendation on the Ethics of Artificial Intelligence, adopted by 193 member states in November 2021

(UNESCO, 2021). Adopted on a voluntary basis, this rights-based framework is founded on the following values:

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity;
- Environment and ecosystem flourishing;
- Ensuring diversity and inclusiveness; and
- Living in [a] peaceful, just and interconnected society.

It is based on the following principles:

- Proportionality and do no harm;
- Safety and security;
- Fairness and non-discrimination;
- Sustainability;
- Right to privacy and data protection;
- Human oversight and determination;
- Transparency and explainability;
- Responsibility and accountability;
- Awareness and literacy; and
- Multi-stakeholder and adaptive governance and collaboration.

It adopts 11 policy areas, including advocating for member states to:

- Conduct ethical impact assessments on the benefits and risks of AI systems;
- Ensure ethical AI governance and stewardship mechanisms;
- Develop data governance strategies to continually evaluate data privacy in AI systems;
- Develop international cooperation on AI-related ethical issues;
- Assess the environmental impact of the life cycle of AI systems; and
- Ensure that AI systems contribute to gender equality, cultural heritage, education and research, access to information and knowledge, labor markets and the economy, and to health and social well-being.

Separately, the CCDH has developed a STAR Framework for international regulation of large technology companies in consultation with regulators, legislators, civil society and academics across the UK, US, EU, Canada, Australia and New Zealand (CCDH, 2023b). The STAR Framework promotes:

- Safety by Design
- Transparency of algorithms, rules enforcement and economics (advertising)
- Accountability to independent and democratic bodies
- Responsibility of companies and their senior executives.

The CCDH suggests that the principle of safety by design should be applied to combat AI-related disinformation. They propose:

- Curating AI training datasets to remove harmful, misleading, or hateful content;
- Employing subject matter experts when developing AI training datasets;
- Constraining AI model outputs to prevent the generation of harmful content; and
- Implementing mechanisms to correct any errors that arise. (CCDH, 2023b)

There also exist private stakeholders invested in ethical AI. The US-based Ethical AI Database (EAIDB) publishes a publicly available, vetted database of AI startups providing ethical services as well as quarterly reports on the state of the ethical AI ecosystem. EAIDB (Raghunathan, 2022a; Raghunathan, 2022b) divides existing ethical AI startups into five main business categories:

1. Data for AI
    a. Startups specializing in the ethical treatment and handling of data for AI, especially regarding issues of privacy, bias and observability.

b. Companies in this category specialize in one of three services:
   i. Data sourcing/observability: ensuring data is sourced properly, e.g. via novel collection methods or ethical labeling solutions and/or offering tools to identify data biases.
   ii. Synthetic data: generating synthetic data, e.g. via statistical methods, to create entirely new synthetic datasets mimicking existing datasets but not linked to real individuals.
   iii. Data privacy: offering cybersecurity services such as data transfer, permissions and control.

2. ModelOps, Monitoring and Observability
   a. Startups in this category monitor models, detect bias and unfairness, mitigate risks, or provide model governance and stakeholder access to bridge the gap between legal, business, and data science teams.
   b. Companies in this category include Fiddler, Arthur, ETIQ, and KOSA.
3. AI Audits, Governance, Risk and Compliance
   a. Specialist consulting firms or platforms that establish accountability and governance, quantify model and/or business risk, or simplify compliance for internal teams within AI systems.
   b. Consulting firms in this category may differentiate themselves by experience or specialization.
4. Targeted AI Solutions and Technologies
   a. These companies build through entire use cases for customers within a single vertical.
   b. Examples include FairPlay AI (fair lending), Pave (fair wage benchmarking) and Zelros (fair insurance recommendation engine).
5. Open-Sourced Solutions
   a. Fully open-source solutions meant to provide easy access to ethical technologies and responsible AI, usually offered by not-for-profit companies.
   b. Most open-source tools are concerned with privacy, bias detection and explainability but face general open-source shortcomings: vulnerability to malicious users, lack of user-friendliness, and lack of extensive support systems.

As another example, the AI safety and research company Anthropic focuses on developing “frontier AI systems that are reliable, interpretable, and steerable” (Anthropic, n.d., “Anthropic”). Anthropic adopts ‘constitutional training,’ which begins with a list of human-created principles, then uses supervised learning and reinforcement learning to build AI systems that are Helpful, Honest and Harmless (HHH) (Bai et al., 2022a). Their HHH framework focuses on creating AI systems that help users, that shares information the system believes to be true while avoiding made-up information, and does not aid the user in harmful activities (Bai et al., 2022b).

However, ethical AI teams in large technology firms have faced layoffs in the first quarter of 2023, when generative AI tools such as ChatGPT began to gain traction. Companies including Microsoft, Meta, Google, Amazon, Twitter and Twitch have reduced staffing on responsible AI teams that advise on the safety of AI-related consumer products (Criddle & Murgia, 2023; De Vynk & Oremus, 2023). While the number of staff affected represent a small fraction of workers affected by layoffs amid a broader tech industry downturn, the timing of the cuts is worrying given the rapid adoption of ChatGPT and other generative AI tools among the general public.

Issues in ChatGPT Regulation

Research-Based Recommendations on ChatGPT

General Recommendations

Hacker et al. (2023) propose that regulation pertaining to generative AI should divide responsibility among four parties across the AI value chain:

1. Developers: those creating and pre-training the model, e.g. OpenAI or Google;
2. Deployers: those fine-tuning the model for a specific use case, e.g. a professional user, though developers and deployers may be the same party;
3. Users: those generating output from generative AI via prompts and putting it into use, either professional or non-professional; and
4. Recipients: those passively receiving and/or consuming the output of generative AI.

They argue that developers should be subject to non-discrimination law and data governance provisions, but that the majority of regulatory obligations should lie with deployers and users regarding risk management systems and performance and robustness thresholds. The obligation for users to implement human oversight and screening of AI systems for evident cases of significant harm should extend to all users, including non-professional users.

As noted above, Huang & Siddarth (2023) highlight that generative AI may degrade the digital commons by flooding online spaces with low-quality content. They suggest three measures for developers and deployers to improve training data quality, which may improve the quality of AI output:

1. Creating consortia to develop best practices around generative AI, with duties such as monitoring, auditing, standards-setting (e.g. on issues such as transparency), and developing shared tools.
2. Establishing ‘gold standard’ datasets for model training or deployment, e.g. by encouraging companies to admit data to privacy-preserving but accessible repositories, which would allow researchers to determine impacts of AI model deployment. This move would also enable companies to adhere to shared standards, but would require establishing transparent data collection procedures as well as external scrutiny on the validity of collected data.
3. Encouraging human feedback to improve models, for instance in return for ownership stakes or governance rights. Such feedback could come from users, experts or specific organizations (e.g. healthcare companies overseeing highly-private data or artists whose works are used in training image-generation models) who interface between data-owners and AI companies. Data-owners may also co-create and co-govern generative AI models themselves: for instance, a group of artists may jointly train and govern an image model that they then monetize.

ChatGPT Regulation: Focus on EU Context

Hacker et al. (2023) argue that EU regulation is “ill-prepared” for generative AI models such as ChatGPT. They highlight four key concerns: 1) data privacy violations under GDPR; and 2) content moderation loopholes under DSA; 3) overly-stringent requirements under proposed AI regulations; and 4) exploitable loopholes under proposed AI regulations.

GDPR: Data Privacy Violations | Generative AI models such as ChatGPT have been shown to be vulnerable to malicious ‘inversion attacks,’ meaning that training data, which may include private information, may be extrapolated from the model. Moreover, GDPR requires information to be provided to users on the processing of personal data provided within a chat interface, especially for minors, but OpenAI has thus far failed to appropriately disclose such information. LLM hallucinations may also contravene GDPR requirements for personal data to be accurate and non-discriminatory.

DSA: Content Moderation Loopholes | The DSA is targeted at illegal content on social networks, not generative AI. The DSA covers intermediary conduits and caching or hosting services, but LLMs do not appear to fall into these categories. Moreover, while DSA may be applied to LLM-generated posts published on social networks, it does not apply to private messaging services such as WhatsApp and Telegram, although problematic content may proliferate in such closed groups. Malicious actors may exploit this loophole to disseminate illegal content at scale, especially with the aid of LLM-generated code.

AI Act: Overly Stringent Regulations | The amendment to the draft AI Act on "general-purpose AI systems" circulated by the French Council Presidency on 13 May 2022 appears overly-stringent in its requirements for developers of generative AI to conduct comprehensive risk management assessments and build risk management systems, given the wide versatility of such AI models. The prohibitive costs of compliance may spur anti-competitive behavior in the generative AI market as only large technology companies may have the resources to comply with these requirements.

AI Act: Exploitable Loopholes | On 7 February 2023, the European Parliament proposed applying even more stringent regulations by classifying all generative AI systems as high-risk, except if the output was subject to human review and if a human or organization was legally responsible for it. However, malicious actors may exploit this exception by subjecting illegal output to human 'rubber stamp' review without changing the content.

In response, they propose four policies for generative AI regulation:

1. Transparency obligations
    a. For developers and deployers: requirements to report on performance metrics and incidents and mitigation strategies on harmful content.
    b. For professional users: requirements to disclose when publicly-available content is generated by or adapted from generative AI.
    c. For non-professional users: implementation of technical measures, e.g. digital rights management, automatic watermarking, and AI-content detection systems, to separate AI-generated content from human-generated content.
2. Risk management via staged release
    a. Adopting staged release for powerful AI models, allowing early access only for security researchers and selected stakeholders to conduct community-based risk assessment ahead of full public release.
3. Non-discrimination strategies in training data
    a. Implementing data curation for representativeness and approximate balance between protected groups at the development and deployment stage.
4. Expanded content moderation
    a. Integrating existing DSA strategies, e.g. mandatory notice and action mechanisms, trusted flaggers, and comprehensive audits for models with many users, into generative AI.
    b. Allowing users to flag problematic content and give notice, with special status accorded to 'trusted flaggers' such as private individuals, NGOs, or volunteer coders, who function as a decentralized content monitoring team.
    c. Obliging content moderation teams working with developers and deployers to respond to notices by modifying the AI system or blocking its output and to establish a comprehensive compliance system for sufficiently-large AI models.

Other Proposed Initiatives: SafeGPT Case Study

Giskard AI, a responsible AI startup, has developed a 'SafeGPT' product targeting LLM issues such as hallucinations, data privacy, toxicity and robustness (accuracy). The product comprises a browser extension for users to identify wrong answers, reliability issues and ethical biases, as well as a quality assurance platform for developers to track performance and debug LLMs by creating business-specific tests. SafeGPT is currently alpha-version and waitlist-only. (Giskard AI, 2023)

International Regulatory Reactions

An opinion piece in The Economist (2023) argues that AI regulation must balance the promises and risks of AI while remaining able to adapt to emerging issues. The piece highlights approaches by different governments, ranging from the UK's "light-touch" approach which extends existing regulations to AI systems, to the EU's "tougher" proposed laws which categorize AI uses by degrees of risk, to China's "sterner" system of dedicated regulation, testing, and approval before public release (The Economist, 2023). The piece suggests that the EU's approach is "closest to the mark" but that a principles-based approach requiring disclosures and inspections "would be more flexible" while allowing greater regulation over time if needed (The Economist, 2023).

Europe

The rapid adoption of ChatGPT led to modifications of the proposed EU AI Act, which lawmakers began drafting nearly two years ago. Journalists report that "a rare example of consensus" among EU politicians concerning the risks of AI technology, with proposed regulations hammered out over just 11 days. The draft act proposes to classify AI tools according to their perceived risk level, from minimal to limited, high, or unacceptable, while areas for concern may include biometric surveillance, misinformation and toxic speech. Lawmakers will likely focus on enforcing transparency around the use of high-risk tools rather than banning such tools. (Coulter & Mukherjee, 2023)

Key concerns within the EU and European countries include:

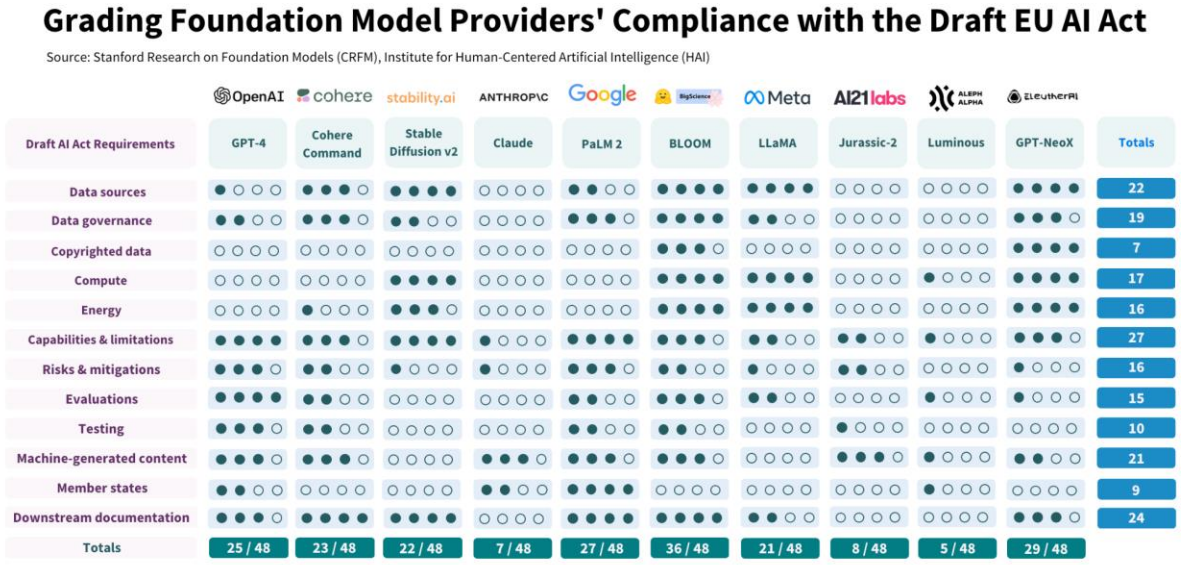
EU: Copyright | The latest draft of the EU AI Act identifies copyright protection as a key issue. Companies deploying generative AI, including ChatGPT, must disclose any copyrighted material used to develop their models (Mukherjee et al., 2023).

Italy: Privacy | On 30 March 2023, Italian Data Protection Authority Garante ordered OpenAI to stop processing personal information from Italian users amid allegations that some users had their messages and payment information exposed to other users, leading OpenAI to take ChatGPT offline on 31 March. On 28 April, access to ChatGPT in Italy was restored amid a raft of measures, including OpenAI adding information on its website on its GPT data collection and training procedures, providing EU users with a form for objecting to using their data for training, and adding an age verification tool for users at the signup stage. (Chan, 2023; Reuters, 2023)

France: Plagiarism | French university Sciences Po became one of the first higher education institutions to ban the use of ChatGPT without explicit referencing by students and faculty, with penalties including expulsion from the university or a ban from French higher education (Kane, 2023; Sciences Po, 2023).

A recent evaluation by the Stanford Center for Research on Foundation Models and Institute for Human-Centered Artificial Intelligence evaluating 10 major foundation model providers finds that providers "largely do not" comply with the current draft of the EU AI Act

(Bommasani et al., 2023). Bommasani et al. (2023) find “a striking range in compliance across model providers” with “significant margin for improvement” even among the highest-scoring foundation model providers, highlighting disclosures of data transparency, copyright, energy use, risk mitigation, and performance auditing standards as consistent challenges (Figure 1). Moreover, they suggest it is “currently feasible” for providers to comply with the AI Act, indicating that the Act could potentially “yield significant change to the ecosystem… towards more transparency and accountability” (Bommasani et al., 2023).



Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	BigScience	Meta	AI21labs	ALEPH ALPHA	EleutherAI	
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	●○○○	●●●○	●●●●	○○○○	●●○○	●●●●	●●●●	○○○○	○○○○	●●●●	22
Data governance	●●○○	●●●○	●●○○	○○○○	●●●○	●●●●	●●○○	○○○○	○○○○	●●●○	19
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●○	○○○○	○○○○	○○○○	●●●●	7
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●	17
Energy	○○○○	●○○○	●●●○	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●	16
Capabilities & limitations	●●●●	●●●○	●●●●	●○○○	●●●●	●●●○	●●○○	●●○○	●○○○	●●●○	27
Risks & mitigations	●●●○	●●○○	●○○○	●○○○	●●●○	●●○○	●○○○	●●○○	○○○○	●○○○	16
Evaluations	●●●●	●●○○	○○○○	○○○○	●●○○	●●●○	●●○○	○○○○	●○○○	●○○○	15
Testing	●●●○	●●○○	○○○○	○○○○	●●○○	●●○○	○○○○	●○○○	○○○○	○○○○	10
Machine-generated content	●●●○	●●●○	○○○○	●●●○	●●●○	●●●○	○○○○	●●●○	●○○○	●●○○	21
Member states	●●○○	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	○○○○	9
Downstream documentation	●●●○	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

Figure 1. Reproduced from “Do Foundation Model Providers Comply with the Draft EU AI Act?”, by R. Bommasani, K. Klyman, D. Zhang & P. Liang, June 15, 2023, Stanford CRFM (crfm.stanford.edu/2023/06/15/eu-ai-act.html). CC BY-NC.

UK

A March 2023 white paper by the UK Department for Science, Innovation and Technology and Office for Artificial Intelligence set out a “pro-innovation” approach to AI (Clarke, 2023). However, UK Prime Minister Rishi Sunak has since underscored the importance of “guardrails” and promoted the UK as “not just the intellectual home, but the geographical home of global AI safety regulation,” announcing in June 2023 that Google DeepMind, OpenAI and Anthropic “have agreed to open up their AI models to the UK government for research and safety purposes” (Clarke, 2023). In June 2023, the UK Department for Science, Innovation and Technology announced the launch of the Foundation Model Taskforce, led by technology entrepreneur, investor and AI specialist Ian Hogarth and backed with an initial £100 million in government funding, to “lead vital AI safety research” (Department for Science, Innovation and Technology).

US

Following US President Joe Biden’s call for companies to ensure AI product safety ahead of release (Miller, 2023), seven AI companies, including OpenAI, Google, Amazon, and Meta, have made “voluntary commitments” to working with the White House on AI safety practices (Siddiqui & Seetharaman, 2023). Separately, Noveck (2023) reports that Boston’s civil service is encouraging a ‘responsible experimentation’ approach to generative AI following a city-wide policy brief sent by City of Boston Chief Information Officer Santiago

Garces on 18 May 2023. The brief encourages city officials to use generative AI on memos, letters, and job descriptions; to translate jargon into approachable language for different target audiences; to translate information into other languages; to summarize text and audio; and as a coding assistant (Noveck, 2023). The city has also integrated Bard into its enterprise Google Workspace to enable access to Bard for all public servants (Noveck, 2023). However, the guidelines highlight that "generative AI is a tool" and that "we are responsible for the outcomes of our tools," with public servants encouraged to fact-check all AI-generated content, disclose the use of generative AI, and avoid sharing sensitive or private information in the prompts (Garces, 2023, pp. 1–4).

China

On 11 April 2023, the Cyberspace Administration of China (CAC) issued draft Measures for the Management of Generative Artificial Intelligence Services to govern generative AI services in China. The draft measures come after the implementation of rules on 'deep synthesis' technologies (Provisions on the Administration of Deep Synthesis Internet Information Services) in November 2022, which targeted audio and visual media over deepfake concerns (Webster et al., 2023). The new draft regulations implicitly target providers of text-based generative AI services following the popularity of ChatGPT and the rise of Chinese competitors, but may also target research and development of such services as per Article 2 (Webster et al., 2023). Commentators highlight the "vague," "broad and demanding requirements" of the draft measures, which include:

- Requiring providers of generative AI services to apply to the CAC for a security assessment and file information regarding its use of algorithms (e.g. the name of the service provider, service form, algorithm type, and algorithm self-assessment report) with the CAC before releasing the generative AI service to the public;
- Holding providers responsible for content produced by generative AI products, fulfill personal information protection obligations, and assume legal obligations of "personal information processing entities" (equivalent to "data controllers" under the EU GDPR);
- Requiring providers to filter inappropriate content and to optimize algorithms to prevent the generation of such content within 3 months;
- Requiring providers to enable the use of tagging mechanisms to identify content/video created by generative AI;
- Requiring that training data must not contain content that infringes intellectual property and to only obtain data on personal information with consent from data subjects. (Webster et al., 2023)

Commentator Helen Toner, Director of Strategy and Foundational Research Grants at the Center for Security and Emerging Technology at Georgetown University, underscores that Article 5 of the Chinese draft measures specifies that companies providing access to generative AI via "programmable interfaces," i.e. APIs such as those released by OpenAI and Google, are responsible for all content produced (Webster et al., 2023). This approach, unlike with the draft EU AI Act, would hold the original AI developers "liable for everything, including issues arising from choices the downstream client company makes about app design or how to restrict user behavior," which appears unfeasible (Webster et al., 2023).

Commentator Paul Triolo, Senior Associate and Trustee Chair in Chinese Business and Economics at the Center for Strategic and International Studies, notes that generative AI chatbots represent a major issue for China's current keyword-based censorship approach, as individual users become "able to ask questions to a generative AI application without any ability to monitor and block the output for sensitivity and offending words" (Webster et al., 2023).

Singapore

Regulatory responses in Singapore have encouraged the ethical use of ChatGPT for schoolwork and business (Ali & Ong, 2023; Chan & Chin, 2023). Singaporean Minister of Education Chan Chun Sing announced in February 2023 that the Ministry of Education was “guiding teachers in schools and institutes of higher learning” on the use of AI tools such as ChatGPT to “enhance learning,” highlighting that students would be taught to critically assess responses from AI tools for accuracy and objectivity as well as to truthfully declare their sources of information (Abdullah, 2023). Journalists have reported that some Singaporean workplaces have been “quick to embrace” the use of ChatGPT at work, especially for idea generation and copywriting (Ali & Ong, 2023).

Separately, Pair, a government-based team, has built a ChatGPT- and Microsoft Word-based service for Singaporean civil servants to increase writing productivity. The tool is expected to help civil servants summarize reference material, explore related ideas, or improve clarity and will be rolled out progressively across agencies to support up to 90,000 civil service. The Government of Singapore has struck an agreement with OpenAI to ensure data privacy around government information. (Chia, 2023)

Conclusion

ChatGPT and other chatbots have proven to be versatile tools offering a range of potential applications, from healthcare to education and art. However, the incidents and experiments outlined in this paper and summarized below (Table 1) demonstrate the urgent need to develop a framework for responsible use of LLMs and generative AI models in general. The threat of disinformation, in particular, remains an especially powerful concern.

As this paper has shown, several frameworks for mitigating the potential threats posed by LLMs and generative AI at large exist, proposed by organizations ranging from nonprofits to private startups as well as researchers. A range of international policy approaches to generative AI exist as well.

Data Governance Concerns	Content Concerns	Social and Economic Concerns	Environmental Concerns
• Data transparency • Data privacy • Cybersecurity (model training and deployment) • Linguistic equity (model training and deployment) • Protection of IP rights • Respect for digital commons	False Content and Disinformation • LLM hallucinations • LLM-generated misinformation and disinformation • LLM use by malicious actors Biased and Dangerous Content • Biased or discriminatory output • Toxic speech • Incitement to dangerous or	Human Labor Replacement Especially among: • Experts • Creatives • Knowledge workers Education, Job Seeking and Hiring • Plagiarism and cheating • Generating expert answers	• Carbon footprint (model training and deployment) • Water footprint (model training and deployment)

	criminal action - Language-based bias (i.e. differences in LLM output based on prompt language)		

Table 1. Summary of concerns across ChatGPT incidents and experiments.

References

Abdullah, Z. (2023, February 12). Students, teachers will learn to properly use tools like ChatGPT: Chan Chun Sing. *The Straits Times*. https://www.straitstimes.com/singapore/politics/students-teachers-will-learn-to-properly-use-tools-like-chatgpt-chan-chun-sing.

Alba, D., & Love, J. (2023, April 19). Google's Rush to Win in AI Led to Ethical Lapses, Employees Say. *Bloomberg*. https://www.bloomberg.com/news/features/2023-04-19/google-bard-ai-chatbot-raises-ethical-concerns-from-employees.

Alba, D. (2023, May 2). AI Chatbots Have Been Used to Create Dozens of News Content Farms. *Bloomberg*. https://www.bloomberg.com/news/articles/2023-05-01/ai-chatbots-have-been-used-to-create-dozens-of-news-content-farms.

Ali, N. H. M, & Ong, J. (2023, March 16). ChatGPT at work: Some S'pore bosses, workers embrace AI tool, others wary or don't admit use to avoid looking 'incompetent.' *Today Online*. https://www.todayonline.com/singapore/spore-bosses-staff-embrace-chatgpt-others-wary-incompetent-2130976.

Ali, R., Tang, O. Y., Connolly, I. D., Zadnik Sullivan, P. L., Shin, J. H., Fridley, J. S., ... & Telfeian, A. E. (2023). Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. *medRxiv*, 2023-03.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2), e35179. doi:10.7759/cureus.35179.

Al-Sibai, N. (2023, February 1). ChatGPT Is Freakishly Good at Spitting Out Misinformation on Purpose. *Futurism*. https://futurism.com/the-byte/chatgpt-minsinformation-newsguard.

Anthropic. (n.d.) *Anthropic*. https://www.anthropic.com/.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022b). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022a). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

Bommasani, R., K. Klyman, D. Zhang & P. Liang. (2023, June 15). "Do Foundation Model Providers Comply with the Draft EU AI Act?". *Stanford CRFM*. https://crfm.stanford.edu/2023/06/15/eu-ai-act.html.

Bordt, S., & von Luxburg, U. (2023). ChatGPT participates in a computer science exam. *arXiv preprint arXiv:2303.09461*.

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494*.

Brewster, J., Arvanitis, L., & Sadeghi, M. (2023, April 24). The Next Great Misinformation Superspreader: How ChatGPT Could Spread Toxic Misinformation At Unprecedented Scale. *NewsGuard*. https://www.newsguardtech.com/misinformation-monitor/jan-2023/.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Campello de Souza, B., Serrano de Andrade Neto, A., & Roazzi, A. (2023). Are the new AIs smart enough to steal your job? IQ scores for ChatGPT, Microsoft Bing, Google Bard and Quora Poe. *Available at SSRN*. https://ssrn.com/abstract=4412505.

CCDH. (2023a, April 5). *Misinformation on Bard, Google's new AI chat — Center for Countering Digital Hate | CCDH*. Center for Countering Digital Hate | CCDH. https://counterhate.com/research/misinformation-on-bard-google-ai-chat/#about.

CCDH (2023b, January 11). STAR Framework: Safety by Design. *Center for Countering Digital Hate | CCDH*. https://counterhate.com/blog/star-framework-safety-by-design/.

Chan, G., & Chin, M. (2023, March 2). No ban on students using AI tool ChatGPT for schoolwork, but ethical use will be taught: IB. *The Straits Times*. https://www.straitstimes.com/singapore/no-ban-on-students-using-ai-tool-chatgpt-for-schoolwork-but-ethical-use-will-be-taught-ib.

Chan, K. (2023, April 28). OpenAI: ChatGPT back in Italy after meeting watchdog demands. *AP*. https://apnews.com/article/chatgpt-openai-data-privacy-italy-b9ab3d12f2b2cfe493237fd2b9675e21.

Chia, O. (2023, February 14). Civil servants to soon use ChatGPT to help with research, speech writing. *The Straits Times*. https://www.straitstimes.com/tech/civil-servants-to-soon-use-chatgpt-to-help-with-research-speech-writing.

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). ChatGPT goes to law school. *Available at SSRN*. https://collimateur.uqam.ca/wp-content/uploads/sites/11/2023/01/SSRN-id4335905.pdf.

Christian, J. (2023, February 4). Amazing "Jailbreak" Bypasses ChatGPT's Ethics Safeguards. *Futurism*. https://futurism.com/amazing-jailbreak-chatgpt.

Clarke, L. (2023, June 12). OpenAI, DeepMind will open up models to UK government. *POLITICO*. https://www.politico.eu/article/openai-deepmind-will-open-up-models-to-uk-government/.

Coulter, M., & Mukherjee, S. (2023, May 1). Exclusive: Behind EU lawmakers' challenge to rein in ChatGPT and generative AI. *Reuters*. https://www.reuters.com/technology/behind-eu-lawmakers-challenge-rein-chatgpt-generative-ai-2023-04-28/.

Department for Science, Innovation and Technology. (2023, June 18). Tech entrepreneur Ian Hogarth to lead UK's AI Foundation Model Taskforce. *GOV.UK*. https://www.gov.uk/government/news/tech-entrepreneur-ian-hogarth-to-lead-uks-ai-foundation-model-taskforce.

De Vynck, G., & Oremus, W. (2023, March 30). As AI booms, tech firms are laying off their ethicists. *Washington Post*. https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International,* 1-15.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 1-15.

Franceschi-Bicchierai, L. (2023, April 20). Jailbreak tricks Discord's new chatbot into sharing napalm and meth instructions. *TechCrunch*. https://techcrunch.com/2023/04/20/jailbreak-tricks-discords-new-chatbot-into-sharing-napalm-and-meth-instructions/.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*, 2022-12.

Garces, S. (2023). *City of Boston Interim Guidelines for Using Generative AI*. City of Boston. https://drive.google.com/drive/folders/1AeTFKh64zWYlhZBZEpNKj8R4dt6ZvW36.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, *9*(1), e45312.

Giskard AI. (2023). *Introducing SafeGPT*. https://www.giskard.ai/safegpt.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246*.

Goren, G. (2023, March 1). Generative AI in HR: ChatGPT, Hiring and Why Career Development is the New Talent Acquisition. *Cangrade*. https://www.cangrade.com/blog/talent-management/generative-ai-in-hr-chatgpt-hiring-and-why-career-development-is-the-new-talent-acquisition/.

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

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative ai models. *arXiv preprint arXiv:2302.02337*.

Haensch, A. C., Ball, S., Herklotz, M., & Kreuter, F. (2023). Seeing ChatGPT Through Students' Eyes: An Analysis of TikTok Data. *arXiv preprint arXiv:2303.05349*.

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., & Liang, P. (2023). Foundation Models and Fair Use. *arXiv preprint arXiv:2303.15715*.

Hosseini, M., Gao, C. A., Liebovitz, D. M., Carvalho, A. M., Ahmad, F. S., Luo, Y., ... & Kho, A. (2023). An exploratory survey about using ChatGPT in education, healthcare, and research. *medRxiv*, 2023-03.

Hsiao, S. (2023, May 10). What’s ahead for Bard: More global, more visual, more integrated. *Google*. https://blog.google/technology/ai/google-bard-updates-io-2023/.

Hsu, T., & Thompson, S. A. (2023, February 13). Disinformation Researchers Raise Alarms About A.I. Chatbots. *The New York Times*. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.

Hu, Y., Ameer, I., Zuo, X., Peng, X., Zhou, Y., Li, Z., ... & Xu, H. (2023). Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Huang, S., & Siddarth, D. (2023). Generative AI and the Digital Commons. *arXiv preprint arXiv:2303.11074*.

Jiao, W. X., Wang, W. X., Huang, J. T., Wang, X., & Tu, Z. P. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Jiao, W. X., Wang, W. X., Huang, J. T., Wang, X., & Tu, Z. P. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Kane, M. (2023, March 20). « J'ai l'impression de vivre une révolution ». À Sciences Po, ChatGPT fait débat. *ZDNet France*. https://www.zdnet.fr/actualites/j-ai-l-impression-de-vivre-une-revolution-a-sciences-po-chatgpt-fait-debat-39955764.htm.

Kent, B. (2023, May 13). *Apricot Blog - ChatGPT vs. Bard: A realistic comparison*. https://blog.theapricot.io/posts/chatgpt-vs-bard/.

Khademi, A. (2023). Can ChatGPT and bard generate aligned assessment items? A reliability analysis against human performance. *arXiv preprint arXiv:2304.05372.*

King, M. GPT-4 aligns with the New Liberal Party, while other large language models refuse to answer political questions. *engrXiv preprint*. https://doi.org/10.31224/2974.

Knight, W. (2023, February 6). The Race to Build a ChatGPT-Powered Search Engine. *WIRED*. https://www.wired.com/story/the-race-to-build-a-chatgpt-powered-search-engine/.

Koyuturk, C., Yavari, M., Theophilou, E., Bursic, S., Donabauer, G., Telari, A., ... & Ognibene, D. (2023). Developing Effective Educational Chatbots with ChatGPT prompts: Insights from Preliminary Tests in a Case Study on Social Media Literacy. *arXiv preprint arXiv:2306.10645*.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health*, *2*(2), e0000198.

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

Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., & Eger, S. (2023). ChatGPT: A Meta-Analysis after 2.5 Months. *arXiv preprint arXiv:2302.13795*.

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI Less" Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv preprint arXiv:2304.03271*.

Li, X., Zhang, Y., & Malthouse, E. C. (2023). A Preliminary Study of ChatGPT on News Recommendation: Personalization, Provider Fairness, Fake News. *arXiv preprint arXiv:2306.10702.*

Mbakwe, A. B., Lourentzou, I., Celi, L. A., Mechanic, O. J., & Dagan, A. (2023). ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digital Health*, *2*(2), e0000205. https://doi.org/10.1371/journal.pdig.0000205.

McGee, R. W. (2023). Is ChatGPT biased against conservatives? An empirical study. *Available at SSRN*. https://www.researchgate.net/profile/Robert-Mcgee-5/publication/368621918_Is_Chat_Gpt_Biased_Against_Conservatives_An_Empirical_Study/links/641caf9da1b72772e420b3b4/Is-Chat-Gpt-Biased-Against-Conservatives-An-Empirical-Study.pdf.

McGuffie, K., & Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

Meschuk, J. (2023, March 1). What ChatGPT and other generative AI tools mean for HR. *Fast Company*. https://www.fastcompany.com/90857953/what-chatgpt-and-other-generative-ai-tools-mean-for-hr.

Metz, C., & Isaac, M. (2023, February 7). Meta, Long an AI Leader, Tries Not to Be Left Out of the Boom. *The New York Times*. https://www.nytimes.com/2023/02/07/technology/meta-artificial-intelligence-chatgpt.html.

Metz, R. (2023, April 18). People Are Using AI for Therapy, Even Though ChatGPT Wasn't Built for It. *Bloomberg*. https://www.bloomberg.com/news/articles/2023-04-18/ai-therapy-becomes-new-use-case-for-chatgpt.

Miller, Z. (2023, April 6). Biden says tech companies must ensure AI products are safe. *AP News*. https://apnews.com/article/joe-biden-artificial-intelligence-science-technology-chatgpt-6948df344041ef1e794d199595bf69e9.

Milmo, D. (2023, February 3). ChatGPT reaches 100 million users two months after launch. *The Guardian*. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

Morvan, J. (2023, February 13). Oups, la nouvelle IA de Google balance une fake news. *Konbini*. https://www.konbini.com/internet/oups-la-nouvelle-ia-de-google-balance-des-fake-news/.

Mukherjee, S., Foo, Y. C., & Coulter, M. (2023, April 28). EU proposes new copyright rules for generative AI. *Reuters*. https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/.

Murgia, M., & Criddle, C. (2023, March 29). Big tech companies cut AI ethics staff, raising safety concerns. *Financial Times*. https://www.ft.com/content/26372287-6fb3-457b-9e9c-f722027f36b3.

Murgia, M. (2023, April 14). OpenAI's red team: the experts hired to 'break' ChatGPT. *Financial Times*. https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8.

Navarra, K. (2023, April 5). ChatGPT and HR: A Primer for HR Professionals. *SHRM*. https://www.shrm.org/resourcesandtools/hr-topics/technology/pages/chatgpt-hr-primer-ai-workplace.aspx.

Noveck, B. S. (2023, May 19). Boston Isn't Afraid of Generative AI. *WIRED*. https://www.wired.com/story/boston-generative-ai-policy/.

OpenAI. (2023a). *GPT-4*. https://openai.com/research/gpt-4.

OpenAI. (2023b). *Introducing ChatGPT*. https://openai.com/blog/chatgpt.

Oremus, W. (2023, May 5). He wrote a book on a rare subject. Then a ChatGPT replica appeared on Amazon. *Washington Post.* https://www.washingtonpost.com/technology/2023/05/05/ai-spam-websites-books-chatgpt/.

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

Parisi, K. (2023, March 14). This company is experimenting with using ChatGPT for performance reviews. *HR Brew*. https://www.hr-brew.com/stories/2023/03/14/this-company-is-experimenting-with-using-chatgpt-for-performance-reviews

Passaglia, S. (2023, April 6). How well does ChatGPT speak Japanese? *Sam Passaglia*. https://passaglia.jp/gpt-japanese/.

Pierce, D. (2023, March 21). Testing Google Bard: the chatbot doesn't love me, but it's still pretty weird. *The Verge*. https://www.theverge.com/2023/3/21/23650472/google-bard-ai-chatbot-hands-on-test.

Pringle, E. (2023, March 28). We asked Google's A.I. chatbot 'Bard' basic SAT questions and it would flunk a real exam. *Fortune*. https://fortune.com/2023/03/28/google-chatbot-bard-would-fail-sats-exam/.

Raghunathan, A. (2022a, June 17). The Ethical AI Startup Ecosystem 01: An Overview of Ethical AI Startups. *Montreal AI Ethics Institute.* https://montrealethics.ai/an-overview-of-ethical-ai-startups/.

Raghunathan, A. (2022b, May 14). EAIDB: Ethical AI Ecosystem Database. *Medium*. https://medium.com/@abhinavr2121/the-ethical-ai-ecosystem-market-map-39779a9ea4ce.

ResumeBuilder. (2023, February 13). 3 in 4 Job Seekers Who Used ChatGPT to Write Their Resume Got an Interview. *ResumeBuilder*. https://www.resumebuilder.com/3-in-4-job-seekers-who-used-chatgpt-to-write-their-resume-got-an-interview/.

Reuters. (2023, April 3). Italian minister says country’s ban on ChatGPT is excessive. *Reuters*. https://www.reuters.com/technology/italian-minister-says-countrys-ban-chatgpt-is-excessive-2023-04-02/.

Rozado, D. (2023a). The political biases of ChatGPT. *Social Sciences*, *12*(3), 148.

Rozado, D. (2023b, February 2). The unequal treatment of demographic groups by ChatGPT/OpenAI content moderation system. *David Rozado*. https://davidrozado.substack.com/p/openaicms.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text be Reliably Detected?. *arXiv preprint arXiv:2303.11156*.

Sadeghi, M., & Arvanitis, L. (2023, May 5). Rise of the Newsbots: AI-Generated News Websites Proliferating Online. *NewsGuard*. https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/.

Saenko, K. (2023, May 23). Is generative AI bad for the environment? A computer scientist explains the carbon footprint of ChatGPT and its cousins. *The Conversation.* https://theconversation.com/is-generative-ai-bad-for-the-environment-a-computer-scientist-explains-the-carbon-footprint-of-chatgpt-and-its-cousins-204096.

Schaul, K., Chen, S. Y., & Tiku, N. (2023, April 19). See the websites that make AI bots like ChatGPT sound so smart. *Washington Post*. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

Scheiber, N., & Koblin, J. (2023, April 29). Will a Chatbot Write the Next ‘Succession’? *The New York Times*. https://www.nytimes.com/2023/04/29/business/media/writers-guild-hollywood-ai-chatgpt.html.

Sciences Po. (2023, February 3). Sciences Po bans the use of ChatGPT without transparent referencing. *Espace Presse Sciences Po.* https://newsroom.sciencespo.fr/sciences-po-bans-the-use-of-chatgpt/.

Sedova, K., McNeill, C., Johnson, A., Joshi, A., & Wulkan, I. (2023, March 2). AI and the Future of Disinformation Campaigns. *Center for Security and Emerging Technology*. https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/.

Sejnowski, T. J. (2023). Large language models and the reverse turing test. *Neural computation*, *35*(3), 309-342.

Siddiqui, S., & Seetharaman, D. (2023, July 21). White House says Amazon, Google, Meta, Microsoft agree to AI safeguards. *Wall Street Journal*. https://www.wsj.com/articles/white-house-says-amazon-google-meta-microsoft-agree-to-ai-safeguards-eabe3680.

Smalley, S. (2023). Could ChatGPT supercharge false narratives? *Poynter*. https://www.poynter.org/ifcn/2023/could-chatgpt-supercharge-false-narratives/.

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*.

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An analysis of the automatic bug fixing performance of ChatGPT. *arXiv preprint arXiv:2301.08653.*

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.

Strong, E., DiGiammarino, A., Weng, Y., Basaviah, P., Hosamani, P., Kumar, A., ... & Chen, J. (2023). Performance of ChatGPT on free-response, clinical reasoning exams. *medRxiv*, 2023-03.

Subbaraman, N. (2023, April 28). ChatGPT Will See You Now: Doctors Using AI to Answer Patient Questions. *Wall Street Journal*. https://www.wsj.com/articles/dr-chatgpt-physicians-are-sending-patients-advice-using-ai-945cf60b.

Subhash, V. (2023). Can Large Language Models Change User Preference Adversarially?. *arXiv preprint arXiv:2302.10291*.

Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing*, *7*(1), 35.

The Economist. (2023, May 23). How to worry wisely about artificial intelligence. *The Economist*. https://www.economist.com/leaders/2023/04/20/how-to-worry-wisely-about-artificial-intelligence.

Turley, J. (2023, April 8). Defamed by ChatGPT: My Own Bizarre Experience with Artificiality of "Artificial Intelligence." *Jonathan Turley*. https://jonathanturley.org/2023/04/06/defamed-by-chatgpt-my-own-bizarre-experience-with-artificiality-of-artificial-intelligence/.

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en.

Volokh, E. (2023, March 27). Correction re: ChatGPT-4 Erroneously Reporting Supposed Crimes and Misconduct, Complete with Made-Up Quotes? *Reason*. https://reason.com/volokh/2023/03/22/correction-re-chatgpt-4-erroneously-reporting-supposed-crimes-and-misconduct-complete-with-made-up-quotes/.

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., ... & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2).

Webber, A. (2023). Beamery launches 'world's first' generative AI for HR. *Personnel Today*. https://www.personneltoday.com/hr/beamery-ai-for-hr/.

Webster, G., Toner, H., Haluza, Z., Luo, Y., Dan, X., Sheehan, M., Huang, S., Chen, K., Creemers, R., Triolo, P., & Meinhardt, C. (2023, April 26). How will China's Generative AI Regulations Shape the Future? A DigiChina Forum. *DigiChina*. https://digichina.stanford.edu/work/how-will-chinas-generative-ai-regulations-shape-the-future-a-digichina-forum/.

Zagni, G., & Canetta, T. (2023, April 5). Generative AI marks the beginning of a new era for disinformation. *EDMO*. https://edmo.eu/2023/04/05/generative-ai-marks-the-beginning-of-a-new-era-for-disinformation/.
Zamfiroiu, A., Vasile, D., & Savu, D. (2023). ChatGPT–A Systematic Review of Published Research Papers. *Informatica Economica*, *27*(1), 5-16.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring AI ethics of ChatGPT: A diagnostic analysis. *arXiv preprint arXiv:2301.12867.*