GENERATIVE AI AND DEMOCRACY

JUDITH SIMON

Introduction

Generative AI has taken the world by storm. This most recent summer of AI started with the launch of ChatGPT in November 2022. The usage numbers exploded and within the first two months only ChatGPT had already reached a threshold of 100 million users, a benchmark the most successful social media sites, such as TikTok and Instagram, needed considerably longer to reach.[1] In the years since, Generative AI has grown substantially with new products and services being announced with breathtaking speed. On the one hand, Generative AI is no longer confined to text but is now equally applicable to the generation of pictures as well as audio and video files, with models and tools such as Stable Diffusion, DALL-E, and Gemini already in wide usage, while others, such as the video-generator SORA, have been announced but have not yet been released at the time of writing. On the other hand, such tools are increasingly being integrated into other

1 By comparison, TikTok needed nine months and Instagram two years to reach the same threshold. See, for instance: https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

services, such as search (e.g., Bing) or office suites (e.g., Microsoft Copilot) and organizational processes in various domains. Given the failure to halt or only slow down this process through invocations of moratoria (some credible, most not), the speed of development and uptake is likely to increase rather than decrease. As a result, generative AI continues to affect a wide range of societal domains, causing upheavals in journalism and education, in science, in medicine and psychotherapy, in public administration, and in the justice system.

The core of generative AI is the capacity to produce new verbal or visual products of increasingly high quality based on patterns discovered in massive amounts of data. The difference between generative AI and earlier developments is not only improved performance, but also the fact that the tools are no longer restricted to specific domains. Due to the fundamental role of language and images for human interaction, the capacity to produce text, pictures, or videos on any topic imaginable — with high plausibility but no relation to truth — should not be underestimated: while language is the central medium of human communication, images and videos are of crucial importance for questions of evidence, for testimony, and for memory, as well as for emotions.

Apart from the high quality of output and the breadth of applicability, another important aspect that explains the unprecedented uptake of ChatGPT and other tools making use of Generative AI concerns their very high usability and availability through simple interfaces and free access via the Internet. Users need almost no previous knowledge and only low technical requirements to be able to produce and distribute texts, images, or videos of a very high quality in a matter of seconds. Prompt the systems with any request and a text or picture can be produced and amended in no time and with little effort.

These aspects explain the extremely rapid spread of ChatGPT, Dall-E, and co. — with all the positive and negative consequences associated with these AI systems. Generative AI now has many millions of regular users, billions of requests, and corresponding results, which can be used and abused for a wide variety of purposes. We therefore need to assess and combat real challenges and dangers for democracy while not being distracted by bogus debates. The latter includes, for instance,

debates surrounding singularity and the end of humanity, as well as the somewhat misguided discussion about whether ChatGPT shows signs of general artificial intelligence, real understanding, or even consciousness. To be very clear: none of the current AI systems has any true understanding of the output it produces, let alone consciousness. ChatGPT detects speech patterns, the probability of word combinations, and the linguistic structure of different genres of text, based on an analysis of vast amounts of text, and it produces new texts based on these learned patterns. While one may argue that such a recognition of linguistic patterns is necessary to human understanding and that multi-modal AI systems linking text to pictures may indeed even approach a next level of "understanding", it appears farfetched to say that this is understanding in a full sense. Thus, even if ChatGPT and co. may *appear* to understand us when answering our prompts, it is worth repeating that the output generated is purely based upon the statistical analysis and reproduction of text without any true understanding of the content.

The Problems of Deception

This appearance, however, indicates one central problem of generative AI: that of deception. Indeed, Generative AI creates at least four different problems of deception.[2]

First, there are potential dangers when users wrongly believe that they are interacting with a human rather than with a machine, e.g., in contexts such as customer service or — much more problematically — in therapeutic contexts. Apart from this most obvious problem of deception, there is also the problem of deception regarding the *capabilities* of AI. Although current AI systems have neither understanding nor consciousness, it can *appear* to users that they do — even if users *know* that they are interacting with a machine. This inclination was demonstrated by early users of Weizenbaum's (1966) natural language processing software ELIZA, and is reflected in current reports on user interactions with ChatGPT. It is sometimes difficult to discern whether people truly

2 Please see Simon (2024/forthcoming) for a full-fledged analysis of the four kinds of deception caused by Generative AI.

believe that ChatGPT, Lambda, and other AI-based chatbots understand them or are conscious, or whether they are merely intentionally feeding the AI hype cycle. However, such an attribution of abilities to a machine by a human says little to nothing about the machine, but a lot about the human tendency to anthropomorphize technology. Indeed, this conflation between the *performance of "speech"* and the *ability to think* is inherent in the discourse on artificial intelligence from its inception, tracing back to the Turing Test (1950) and Searle's (1981) critique thereof.[3]

By highlighting the difference between the (nonexistent) competence of machines and the ways humans are deceived by these machines' performance, I do not aim to scoff at this human error. On the contrary, I want to issue a warning about the *performative power of simulation*: simulating intelligence, understanding, or even emotion and empathy, even if it is only a simulation, has real implications and makes us as humans vulnerable. We react cognitively and emotionally to language and images in a special way — and that is what makes these new technologies simultaneously immensely powerful and potentially harmful.

The third form of deception then concerns the deceptive results these systems generate, from funny pictures of Pope Francis in unusual clothing and videos "resurrecting" historic figures, to revenge porn, fake news, and deepfakes for propaganda purposes, from audio files "reviving" loved ones, to the criminal use of fake voices to deceive relatives. This last problem in particular poses serious challenges to societal communication and the stability of democracies. Of course, deception, propaganda, and manipulation are not new subjects. But the ease and speed with which high-quality texts, images, sound files, and videos can now be produced and distributed in real time through social media and messenger services

3 With the so-called Turing Test (1950), Alan Turing suggested that when a human cannot distinguish whether the responses to her queries are coming from a machine or from another human then this would be a sign of machine intelligence. John Searle countered this conclusion with his famous *Chinese room argument* (1981) in which he argued that merely successfully manipulating Chinese symbols by executing linguistic rules can and thus should be distinguished from understanding Chinese, i.e., the meaning of such symbols.

opens up a completely new dimension of possible misuse. By flooding the public sphere with false but plausible looking content, generative AI tools pose a real danger to our democracies, as fundamental processes of information and communication can be disrupted quickly, easily, and with potentially severe and lasting impacts.

The fourth and final form of deception concerns problems resulting from the integration of Generative AI into other services and products, such as web search, email programs, and office suites. Indeed, ChatGPT was initially heralded as the future of search. This misleading depiction of functions and underlying processes causes a conflation of information retrieval with information creation, which poses further challenges for evaluating information.

Taken together, these four different types of deception can cause severe epistemic, ethical, and political harm. Deception may not only cause false beliefs; the increasing difficulty of assessing the truthfulness of content may also decrease overall trust in practices and institutions of information themselves. If people feel they cannot reliably judge the quality of information nor the providers, this has potentially severe implications for public communication and democracies.

Where Do We Go from Here?

What do we do now in view of the challenges outlined above and the problems of deception in particular? In my opinion, an appropriate reaction must interweave various instruments. Neither regulation, technology design, nor education are individually sufficient, but jointly they provide our best bet to counter the challenges to democracy posed by Generative AI.

Regulation can be achieved through various forms of hard and soft law. Overall, I am rather skeptical of self-regulation in this context as industry leaders have so far not exhibited much ethical sensibility and the fear of missing out for the companies is simply too great. For that reason, we cannot rely on industry stakeholders to voluntarily commit to control their products, but rather we must ensure sound democratic control and oversight of these technologies.

In the European context, a number of laws addressing AI are already in place or are about to enter into force, such as the General Data Protection Regulation, the Digital Services Act, and the Digital Markets Act. However, the most important law in that context is the EU AI Act, which was passed in 2024. The EU AI Act, initiated before the advent of Generative AI, endorses a risk-based approach in which the regulation of AI depends on the context and sector of application. It therefore proposes specific requirements for the deployment of AI, but only in high-risk sectors or contexts, such as medicine or education. This focus on regulating critical sectors of application rather than regulating AI *tout court* has merits, but Generative AI, or so-called *general-purpose AI*, has more generally demonstrated the limits of this approach as one of its core features is precisely to be applicable *across* different sectors and domains.

So how should we regulate Generative AI then? Only when it is used in critical contexts, thereby placing the main responsibility for regulation on the deployers and professional users of Generative AI? Or do we want to place demands on the producers of Generative AI? During the AI Act Trilogue in December 2023, these questions were hotly debated within and between the European Parliament, the European Council, and the European Commission. In the end, an agreement was reached to amend the previous proposal of the AI Act and to include rules about high-impact general-purpose AI models that can cause systemic risk in the future. As the AI Act has yet to enter into force, both the interpretation of the law and its effectiveness are still open. And so is the discussion of how exactly obligations should be distributed fairly and effectively between producers, deployers, and (professional) users of such systems to respect fundamental rights and societal values.

Returning to the problem of deception, various measures for increasing transparency have been proposed. The first obvious solution concerns mandatory labeling of content provided by or with the help of AI, a requirement that is already included in the EU AI Act. Indeed, various technical solutions are currently being developed to either detect fakes or to verify true content through watermarks. Such mandatory labeling as well as technological solutions are important, but they are not sufficient to counteract the problems of deception. This labeling will not

rule out criminal misuse or informational warfare. It also will not prevent people from ascribing properties to technologies that they do not have. Accordingly, there is a need to develop new norms and competencies to deal with these AI systems and to clearly identify the possibilities as well as the limitations of these systems.

More transparency can also be achieved at the level of the models themselves through open access. While ChatGPT provides free access to its basic services without fees, it is otherwise a completely proprietary and opaque system. In contrast, open-source alternatives, such as BLOOM or Stable Diffusion, allow inspecting, testing, and even modifying of the underlying technology. Such openness of course also comes with its own problems as free and open access also enables new forms of abuse. Therefore, it will be necessary to carefully examine which forms of openness have the most advantages and the fewest disadvantages. The case of ChatGPT, which combines easy and free basic access to an otherwise completely opaque and proprietary system, seems like the worst possible combination.

Finally, education is of special importance to Generative AI in at least three regards. First, it is an area that was and continues to be in itself profoundly challenged by Generative AI. Second, it is considered a high-risk area for the deployment of (Generative) AI. And third, education is central to countering the challenges Generative AI poses to democracy.

The rapid uptake of ChatGPT initially presented universities and schools with the challenge of making exams as fraud-proof as possible. The primary question revolved around how to guarantee fairness if some students have ChatGPT do their homework and others do not. More fundamentally, however, ChatGPT also opens up the possibility — and the necessity — of asking ourselves about the nature and value of education. When even students of literary studies delegate the writing of their essays to ChatGPT and texts that appear to be scientific are produced with fabricated sources, then what does this say about the goals of education and the enabling conditions at universities? Which skills and abilities still have to be learned under the condition of new technological possibilities, which ones have to be added, and which ones might no longer be needed? The German Ethics Council has provided

some orientation for answering these questions in the report "Humans and Machines — Challenges Posed by Artificial Intelligence" (Deutscher Ethikrat, 2023). Our core question is how AI can be designed and used in such a way that the possibilities for human agency and authorship of the various actors involved are expanded and not reduced. It appears obvious that education on all levels and in all its different forms must encompass a solid understanding of the nature, premises, and consequences of the technological mediation of our lifeworlds. While this includes scientific, technological, and mathematical knowledge, it does not stop there. Indeed, to make use of the fruits of generative AI, but to avoid being deceived by it, this knowledge needs to be complemented with critical thinking skills, with solid knowledge, expertise and insights from the social sciences, humanities, and the arts. These knowledges crucially also need to be integrated into computer science education, to support the responsible design and development of AI technology from the onset. It is time to reassert that the foundational aim of education is not to provide learners with sellable skills and techniques, but to raise politically mature and responsible citizens. This, in the end, may be one of the biggest challenges we face if we want to secure democratic and sustainable futures.

References

Deutscher Ethikrat. (2023, March 20). *Stellungnahme Mensch und Maschine — Herausforderungen durch Künstliche Intelligenz*. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

Searle, J. (1981). Minds, Brains, and Programs, *Behavioral and Brain Sciences*, 3, 417–57. https://doi.org/10.1017/S0140525X00005756

Simon, J. (2004/forthcoming). Generative AI, Quadruple Deception & Trust, *Social Epistemology, Special Issue: The Mind-Technology Problem: Rethinking Minds, Humans and Artefacts in the Age of AI*.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433–60. https://doi.org/10.1093/mind/LIX.236.433

Weizenbaum, J. (1966). ELIZA-A Computer Program for the Study of Natural Language Communication Between Men and Machines. *Communications of the ACM*, 9, 36–45.