Routledge
Taylor & Francis Group

Social Epistemology

A Journal of Knowledge, Culture and Policy

ISSN: 0269-1728 (Print) 1464-5297 (Online) Journal homepage: www.tandfonline.com/journals/tsep20

Generative AI, Quadruple Deception & Trust

Judith Simon

To cite this article: Judith Simon (08 May 2025): Generative AI, Quadruple Deception & Trust, Social Epistemology, DOI: 10.1080/02691728.2025.2491087

To link to this article: https://doi.org/10.1080/02691728.2025.2491087

Published online: 08 May 2025.

SOCIAL EPISTEMOLOGY
https://doi.org/10.1080/02691728.2025.2491087

Routledge
Taylor & Francis Group

Generative AI, Quadruple Deception & Trust

Judith Simon

Department of Informatics, University of Hamburg, Hamburg, Germany

ABSTRACT

Generative AI has taken the world by storm. With millions of regular users, billions of requests and corresponding results, tools employing Generative AI are ceaselessly used and abused for a wide variety of purposes. This article focuses on the problem of deception resulting from Generative AI and proposes the notion of *quadruple deception* to capture a set of related, yet distinctive forms of deception: 1) deception regarding the ontological status of one's interactional counterpart, 2) deception regarding the capacities of AI, 3) deception through content created with Generative AI as well as 4) deception resulting from integration of Generative AI into other software. Arguing that deception severely challenges practices of assessing trustworthiness and placing trust wisely, I assess the epistemic, ethical and political implications of misplaced trust and distrust resulting from these four kinds of deception. The article concludes with some suggestions on how the trustworthiness of Generative AI could be increased to ground more justified trust and sketches corresponding duties for the design, development and deployment of Generative AI, the discourse about Generative AI, as well as the governance of Generative AI.

ARTICLE HISTORY
Received 1 April 2025
Accepted 6 April 2025

KEYWORDS
Generative AI; deception; trust; trustworthiness

1. Introducing Generative AI

The history of artificial intelligence is one characterized by frequent ups and downs, high hopes and deep disappointments, often described as the summers, respectively winters of AI. Following the launch of ChatGPT in November 2022 and innumerous other tools since then, we are witnessing another summer of AI with Generative AI being at the center of many academic, political and public debates.

There are at least two reasons for this massive interest. First, by making ChatGPT and other tools of Generative AI freely available on the internet, equipping them with a simple and intuitive interface, which made content generation as quick and easy as an online search, uptake exploded.[1] Second, while tools employing techniques and methods from the field of artificial intelligence, in particular machine learning, have been pervading many domains in public and private lives for many years, it was with the advent of ChatGPT that such technologies were also discussed – and perceived – as artificial intelligence. That is, users experienced Generative AI and particularly ChatGPT as intelligent counterparts and not merely as useful software programs.

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 sources, most of them crawled from the web (cf. Bahrini et al. 2023; Bender et al. 2021). The difference to earlier developments in language technologies does not lie in improved performance alone, e.g. fewer grammatical

CONTACT Judith Simon judith.simon@uni-hamburg.de Department of Informatics, University of Hamburg, Vogt-Koelln-Straße 30, Hamburg 22527, Germany

mistakes, the breadth of coverage has also changed as tools are not restricted to specific domains, such as the medical fields, any longer (Bommasani et al. 2022; Qin et al. 2023).

Apart from the high quality of output and the breadth of applicability, another important aspect that explains the unprecedented uptake of ChatGPT and similar tools, concerns their very high usability and availability through simple interfaces and free access via the Internet. The users need almost no previous knowledge and only a few technical prerequisites to be able to produce texts, images or videos in very high quality in a matter of seconds. Prompt the systems with any request, and a text, picture, sound or video file of your liking can be produced and amended in no time and with little effort. These two aspects explain the extremely rapid spread of tools such as ChatGPT, Dall-E or Midjourney – with all the positive and negative consequences associated with it. 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. Moreover, combined with the affordances of social media and messenger services, this content can be shared widely almost in real-time.

Due to the fundamental role of language and images for human interaction, these capacities to produce text, pictures or videos on any topic imaginable – with high plausibility but no relation to truth – cannot be overstated: While language functions as the central medium of human communication, images and videos are of crucial importance for questions of evidence, for testimony, memory but also for eliciting emotions. It therefore appears obvious that developments around large language models and other forms of Generative AI raise numerous epistemological, ethical, and political concerns.

As a form of data-based AI, Generative AI shares all the ethical and epistemological concerns of other types of AI, which have already been widely discussed in academia but also the public and political debate (cf. Deutscher Ethikrat 2023; Mittelstadt et al. 2016; Smuha 2021). These issues can be grouped under a number of different headings such as privacy, data protection and surveillance (Cohen 2008; Nissenbaum 2004); bias, discrimination and fairness (Angwin et al. 2016; Barocas and Selbst 2016; Friedman and Nissenbaum 1997; Simon, Wong, and Rieder 2020); transparency and explainability (Ananny and Crawford 2016); accountability and responsibility (Binns 2017; Lepri et al. 2017); sustainability and working conditions in AI development. Further issues concern transformative effects, the increasing reliance on statistical reasoning (Hacking 1992) in many societal domains may have on autonomy and freedom, but also solidarity and justice (cf. Busch 2016; Deutscher Ethikrat 2023; B. Rieder 2016). All these concerns are pressing and, albeit to varying degrees, characteristic of most if not all data-based AI systems. There are, however, additional concerns which are unique to Generative AI or are at least drastically aggravated through it. These are related to unique interactional features of Generative AI and their propensity to enable different forms of deception.

2. Interacting with Artificial Intelligence as Artificial Intelligence

In recent years, the usage of the term 'Artificial intelligence' has increasingly been equated with machine learning. That is, in contrast to earlier usage of the term which focused on the simulation of behavior which would be considered intelligent, if done by humans, through machines (Minsky 1968), its more recent meaning rather focuses on specific methods of data analysis while not necessarily alluding to the simulation of intelligent behavior any longer. This shift in meaning of the term artificial intelligence appeared to be a dual one: While artificial intelligence research not employing machine learning was increasingly portrayed as outdated, the original goal of AI research to simulate human intelligent behavior with machines appeared less central. As a result, we have been exposed to endless AI-based tools, i.e. tools employing machine learning, but did not necessarily experience these technologies as intelligent. Search engines and recommender software; algorithmic content moderation on social media platform; facial recognitions technologies to unlock our smartphones, to grant or deny access to countries at airports; weather forecasts; and an abundance of tools used to support or even replace human decision making in education, social welfare or the

judicial system – all these tools employ machine learning for data-based classification and prediction.[2] And while these tools clearly simulate certain aspects of intelligent behavior, such as pattern recognition in particular, they are usually not experienced by the users as intelligent counterparts.

This changed with the advent of ChatGPT. Suddenly, the seemingly intelligent machine was back – and it brought back old reactions and debates which have accompanied AI research from its inception. Central to this experience of artificial intelligence as artificial intelligence is the interactional format of ChatGPT as a chatbot, i.e. a software using natural language processing to interact with users as a conversational artificial agent. Thus, while users may use ChatGPT to search for information, not only does it function differently from search engines, it also has a different interactional form. When using a search engine, a user is presented with a multitude of different search results, i.e. content found on the web, indexed, and ranked. ChatGPT differs in two philosophically relevant regards: First, instead of presenting an ordered list of different results and sources, it integrates these into a coherent text. Second, it comes across as a chatbot, i.e. its interface and functioning invites the user to communicate or interact with it by asking questions or making so-called prompts. This simulation of communicative acts can be seen as a return to the original goal of AI to simulate human behavior, in that case, communicative behavior. At the same time, these design features are prone to deceive users into believing, that they interact with a human person instead of a software which merely analyzes and stochastically predicts word patterns.

3. Generative AI and the Danger of Quadruple Deception

The before-mentioned technological features of Generative AI pose various epistemological, ethical, and political challenges related to what I call the threat of quadruple deception.[3] The Oxford English Dictionary defines deception as 'caus[ing] to believe what is false' (Simpson and Edmund 1989, cited from Mahon 2015). Mahon (2015) has argued that this basic definition is too inclusive as it cannot exclude inadvertent and mistaken deception. He thus proposes to define deceiving as 'intentionally caus[ing] to have a false belief that is known or believed to be false' (Mahon 2015). However, for the purpose of this article, I will adopt the broader dictionary understanding of deception, which covers both intended and unintended deception. The reason for doing so is twofold: first, while intent is relevant for the ethical assessment of an action, e.g. its blameworthiness, it is not necessary for detrimental ethical, epistemic, or political consequences of deception to occur. Second, whether or not deception was intended, is often not discernible. I will discuss the merits and potential shortcomings of adopting such a broad notion of deception in section six.

The first form of deception regarding Generative AI concerns the danger that users may be misled into believing that they interact with a human being while indeed interacting with a chatbot. *Deception 1* thus refers to the misconception regarding the *ontological status* of one's counterpart. Contemporary examples for this type of deception may be users who assume they are talking to a customer agent while indeed being confronted with a chatbot or, more worryingly, clients assuming that they interact with a psychotherapist, while they indeed interact with software only.[4] One could argue that with the Turing Test, this type of deception was indeed construed as the benchmark for the realization of artificial intelligence (Turing 1950). While this form of deception is by no means a novel worry, it has become more pressing given the increasing quality and prevalence of tools such as ChatGPT. As a reaction, the proposed AI Act of the European Commission indeed requires that chatbots need to be labeled as such to avoid this form of deception.

The example of the chatbot-psychotherapist leads us to the second type of deception, and once more, into the history of AI: *Deception 2* refers to deception about the *capacities of AI*. Since the launch of ChatGPT and other forms of Generative AI, some have claimed that such tools are more than what Bender et al. (2021) have labeled 'stochastic parrots', but instead supposedly express intelligence, understanding or even consciousness.[5] The tendency of humans to anthropomorphize technologies, i.e. to attribute human characteristics to AI,

accompanies the development of AI from its onset as the classical case of ELIZA illustrates (Weizenbaum 1966). Experiencing users attributing intelligence and empathy to ELIZA – even if they knew it was just a simple software program, turned its creator, Joseph Weizenbaum, into one of the earliest and most fervent critics of AI technologies (Weizenbaum ([1977] 2001)). Such deception about the capacities of AI is currently of very high concern again. And while it is easy to debunk many of these claims as false, they left an impression – on the public and political, but also on academic debates (cf. Coeckelbergh and Gunkel 2023; Nyholm 2023).

The third type of deception related to Generative AI, *Deception 3*, concerns the deception caused by *misleading content* produced with Generative AI. Examples of this type of mis- or disinformation are supposedly scientific publications with faked references created with the help of ChatGPT but also deepfakes in the form of images, videos, or audio files. The potential impact of such content ranges from amusing,[6] to severely damaging, as in the case of the fake audio files used to affect elections in Slovakia or the US.[7] Thus, the impact of such deceiving content used for manipulation and propaganda cannot be overestimated. Clearly, manipulation and propaganda have a long history, and the role of technologies in aggravating these dangers has been extensively addressed and assessed in the public, but also in philosophical analyses (Fallis and Mathiesen 2019; Gelfert 2018; Hendricks and Vestergaard 2019). However, the combination of Generative AI and Social Media has massively increased the threat of this type of deception due to the ease and speed with which misleading content of sufficiently high quality can be produced and disseminated. Indeed, the Global Risk Report 2024 of the World Economic Forum lists AI-generated misinformation and disinformation as the most severe anticipated global risk for the next two years (World Economic Forum 2024).

The fourth type of deception, *Deception 4*, concerns deception regarding the *function* of Generative AI. From its launch, ChatGPT and its competitors were heralded as the future of search engines.[8] And while there are similarities in the usage of ChatGPT and similar tools with search engines, this comparison is deeply misleading. The user may indeed use ChatGT to search for information online and in many cases, the results may look similar to information found online, e.g. on Wikipedia. However, the functioning of a search engine differs from that of an LLM in epistemologically highly significant ways: most importantly, LLMs are not *retrieving* existing texts, but *generatin*g new texts. While this novel stochastically produced text is based upon the materials found online, because the text patterns are extracted from the training material found online, this difference matters epistemically – and it adds an extra layer of potential misinformation and disinformation regarding the epistemic status of this text.

Companies offering search engines may have various motives for using different forms of AI in the context of search, e.g. to personalized content or to increase the usability and user experience for users with different preferences or needs, for instance by offering multi-modal input or output channels for their services. While personalization has its own epistemic, ethical, and political challenges ranging from privacy concerns (Nissenbaum 2011) to debates on filter bubbles and echo chambers (cf. Nguyen 2020; Pariser 2011), I want to point here to a different and more specific concern, namely the epistemic consequences resulting from a blurring of boundaries between retrieving and creating information. Put more bluntly: if AI is not only used to personalize content, but to create personalized content, this results in a dangerous liaison of personalization and deception.[9]

Before relating these four forms of deception to the issue of trust, I want to address one potential counterargument. Of course, there are numerous epistemic problems with content found on the internet, as this may be just as false or misleading as stochastically produced nonsense, i.e. the impact of human-generated lies can have the exact same impact as AI-generated nonsense. And indeed, the epistemic pitfalls of blindly relying on online sources have been addressed in social epistemology and related fields for decades (cf. for instance, Fallis 2000; Goldman 1999, 2008; Miller and Record 2013, 2016; Simon 2010a, 2010b, 2015). My claim here is that the blurring of use

functionalities between information retrieval and information creation poses an *additional* epistemic challenge. As such LLMs are likely to trump and exacerbate the epistemic, ethical and societal harms posed by other digital technologies.

4. Deception, Trust, and Trustworthiness

A fruitful lens to understand the epistemological, ethical and political significance of these four types of deception is trust. The philosophy of trust has emerged as a blooming field of inquiry not only within ethics (Baier 1986), but also in the philosophy of science (Hardwig 1991) as well as political and social philosophy (cf. Simon 2020a). Trust is of practical concern when it comes to relying on other persons both for daily tasks and interactions, as well as of epistemic concern when relying on others for gathering information and acquiring knowledge. Moreover, trust in institutions and organizations plays a fundamental role in organizing collective life. Finally, while it remains controversial whether technologies can be patients of trust themselves, technologies in general, and information and communication technologies in particular, mediate trust relations amongst humans, between humans and organizations as well as between humans and information (cf. Simon 2020b).

Trust is often conceived as desirable. It is *intrinsically* desirable because trusting or being trusted is often considered a value in itself. It is conceived as *instrumentally* valuable, because trust enables other valued goods, such as cooperation. However, from a normative perspective, it has been argued that trust is not valuable per se, but if and only if it is directed at those who are indeed trustworthy, because misplaced trust can lead to exploitation and harm. Thus, it is the relation and fit between trust and trustworthiness that is of utmost philosophical concern (O'Neill 2020; Scheman 2020; Simon 2020b). Indeed, the ethical, epistemological, and political consequences of trust wrongly placed or withheld have been subject to intense philosophical scrutiny in recent years. While I cannot provide more detail on the intricacies of trust and trustworthiness in this short contribution, I sketch in the following how the different types of deception outlined above affect questions of trust and its relation to trustworthiness.[10]

Deception negatively affects trust in various ways. First, I may falsely place trust in someone or something deceptive and may as a result be practically or epistemically harmed, e.g. by acquiring and believing false information or by being let down otherwise. Moreover, deception can affect trust also indirectly after being revealed, i.e. detecting deception diminishes trust. Yet, while it indeed should diminish trust into the actor who deceived me, it may also undermine trust more widely. If feel I cannot distinguish trustworthy from untrustworthy sources any longer, I may, as a consequence of having been deceived, suspend from trusting – or even distrust – other actors irrespective of their trustworthiness. Thus, deception not only diminishes trust in those people or organizations that deceive. The overall atmosphere of trust is threatened by deception if it becomes increasingly difficult to discern trustworthy sources and to distinguish information from mis- and disinformation. As such, deception has direct and indirect negative consequences for assessing trustworthiness and placing trust wisely.

5. Generative AI, Quadruple Deception, and Trust

It has been argued above that Generative AI is prone to different types of deception and that deception undermines trust. But how exactly does deception about Generative AI threaten interpersonal, epistemic, and societal trust? In the following, I sketch the implications the different forms of deception may have on trust. And while I hope to have shown, that distinguishing the different forms of deception is analytically valuable, it will become obvious that they are often related and affect various forms of trust in different, yet entangled ways.

The most straightforward form of deception through Generative AI occurs regarding its ontological status as a counterpart: are you interacting with a chatbot or with a human person? Deception in this context can happen because one simply falsely assumes to be interacting with an employee in

customer service or – even more worryingly, with a psychotherapist – while indeed interacting with software. Such deception can be unintended, but also intended, comparable for instance to fake profiles on social media sites such as X spreading propaganda. This form of deception affects trust in different ways. First, falsely believing to interact with a human while indeed dealing with a chatbot can lead to unjustified trust in the capacities of this counterpart, such as its supposed understanding or empathy, thereby relating *Deception 1* to *Deception 2*, as well as into the information provided by the chatbot, thereby relating it to *Deception 3*. In both cases, the trust placed in the software falsely assumed to be a human is unjustified and can cause ethical, epistemic, and societal harm.

At the intersection of *Deception 1* and *Deception 3*, epistemic trust, in particular, may be challenged if one happens to believe false information provided by manipulative fake profiles on social media sites. This would be a case of unjustified epistemic trust in an untrustworthy source, leading to epistemic and possibly also practical or ethical harm depending on the kind of deception occurring.[11] Yet, not only the deception itself, but also the revelation of the deception regarding the ontological status of one's counterpart affects trust in different ways. Under ideal circumstances, a reduction of trust may lead to a more appropriate level of trust into the chatbot where the expectations of the human user match the real capacities of the chatbot. It may, however, also cause a more widespread and detrimental loss of trust into the institution employing the chatbot – or even into societal communication at large – if one feels that reliably distinguishing between humans and machines becomes difficult or impossible.

Deception 2 refers to the various ways in which AI is being characterized or perceived to have capacities, such as understanding or consciousness, which are usually either reserved for human beings or, at most, extended to other living beings, i.e. to certain animals. The degree to which people truly believe claims about AI being sentient or are merely stating them for strategic purposes is not always clear. However, given the functioning of Generative AI, those software systems can be said to 'understand' only in a rather reductionist sense and certainly do not possess consciousness irrespective of whether their *behavior* may appear to signal the existence of such capacities to the users. This discrepancy between how things are and how they appear, opens the possibility for deception. What would be the implications of *Deception 2* for trust? First of all, there is a danger of misplaced trust in such systems. If one assumes that a chatbot truly understands the meaning of one's own communicative acts instead of merely processing the words and producing a likely response, this elicits certain expectations not only regarding the communicative behavior of such systems but also regarding its *intentional stance* (Dennett 1971) towards the user.

Take the example of chatbots used in therapeutic contexts. When interacting with a human psychotherapist, the client may not only have certain contextual expectations regarding the conversational behavior of the psychotherapist, but also regarding her stance towards him and the relationship that binds them together. And while the communicative behavior of a psychotherapist may be simulated to a reasonable degree by a chatbot, it is also the intentional stance of the psychotherapist and the relation between client and psychotherapist, which is crucially relevant for most therapeutic approaches. By stating that this relational aspect may be relevant for most, but not all psychotherapeutic approaches, I grant that defenders of strictly behaviorist forms of psychotherapy may plausibly deny the necessity of any relationality between client and psychotherapist and thus would also accept a chatbot psychotherapist as a full substitute for a human psychotherapist.[12]

Deception 3 concerns deceptive pictures, videos or text produced with Generative AI. This development, often discussed under the heading of *fake news* or *deep fakes*, is not new. However, Generative AI makes the creation and distribution of fake content incredibly quick and easy. While ChatGPT and other LLMs can be used to create deceptive texts, using Generative AI for creating deceptive visual and auditory content may cause even more severe societal problems, resulting from misplaced trust.

Deceptive content can, first and foremost, lead to unjustified trust in this content and those providing it. Thus, from an epistemological perspective, deepfakes may lead to false beliefs, which may have severe societal consequences. And indeed, this worry does not appear far-fetched as

various incidents of manipulative use of deepfakes have recently been reported in different countries.[13] Moreover, the difficulty to distinguish between true and false content can lead to an overall reduction of trust within societies. If fake content is presented as scientific evidence, if faked statements by politicians are circulated, overall trust in politics, the media, or science can be diminished as a result, irrespective of their actual trustworthiness. Accordingly, faked content created with the help of Generative AI poses significant societal challenges by soliciting trust in untrustworthy content and sources as well as by making the distinction between false and true information more difficult to impossible, thereby eroding trust overall and thus also trust in trustworthy sources. As such, the negative implications of deep fakes through Generative AI are a major threat for contemporary democracies by severely and negatively affecting epistemic, interpersonal, and societal trust.

Finally, both the presentation of tools such as ChatGPT as the future of search[14] as well as the integration of Generative AI into various tools of information retrieval also poses significant challenges for epistemic and ethical trust. Information retrieval, e.g. in the form of online search but also search within one's email program works, under the premise that *existing* content is being searched for. Generative AI, however, as the term 'generative' indicates, is not about finding, but creating content based upon patterns in the data upon which it was trained. As such, the integration of Generative AI into search and other services is a case of 'function creep' (Koops 2021) with significant consequences for epistemic trust in particular. After all, how can I rely upon and trust these tools – or my memory – if I cannot be certain that the email I am 'finding' existed before and was hopefully also written by a human being – and not created *as a consequence* of my search? Clearly, one may hope that the use of Generative AI is mostly restricted to personalization or multi-modal content provision, but how can we be sure if we do not understand the underlying functioning of the newly emerging AI button suddenly appearing in different types of service ranging from email and office software packages to search engines and pdf readers.[15] The consequences of this integration of Generative AI and the resulting con/fusion of functionalities may have on epistemic and cognitive processes are far from sufficiently addressed or even assessed. Surely, we will learn how to deal with these novel integrated systems. Yet given the speed of development and integration of Generative AI and the lack of knowledge and understanding regarding the inner workings of these systems, negative implications for trust in other persons, institutions or content appear plausible, as well as for trusting oneself and one's memory.

6. Countering a Counterargument: Are Deception 1–4 Really Instances of Deception?

Before concluding and providing some suggestions on how to mitigate the challenges related to the four types of deception outlined above, let me address a fundamental challenge to my line of reasoning, namely, whether the four issues outlined above are instances of deception. The possibility of classifying the challenges above as cases of deception depends upon one's definition of deception: if deception requires intention to deceive (cf. Mahon 2015), one may challenge whether all cases of deception outlined above are really intended by the developers or deployers of AI or whether not at least some of them are not intended and should thus rather be conceived as cases of misjudgments or mistakes by the users. In answering the question whether one can and should characterize all cases described above as deception, one must provide reasons for adopting a wider or more narrow definition of deception and assess the benefits and shortcomings of each choice.

I have provided two reasons for adopting a broader notion of deception, which does not require intent. First, whether or not deception was intended is often not discernible. Second, some of the *consequences* of deception are independent of intent, i.e. 'caus[ing] to believe what is false' (OED 1989, cited from Mahon 2015) can have the *same ethical, epistemic, or political consequences* irrespective of whether the provider of information intended to deceive or not.

To argue why one could and indeed should adopt a broad notion of deception which does not require intent, one may use the case of discrimination as an analogy. The philosophical and legal

discourse on discrimination, which also emerged as a crucial topic within AI ethics, centers around the question whether discrimination fundamentally depends upon the intention to discriminate or whether the differential impact of a given practice on different groups of individuals is sufficient to classify a practice as discriminatory. In the context of the US antidiscrimination law, the two positions correspond to different discrimination doctrines: discrimination as disparate treatment versus disparate impact (cf. Barocas and Selbst 2016). Whereas, in addition to formal disparate treatment of similarly situated people, disparate treatment also requires the intent to discriminate, disparate impact by contrast is not concerned with the intent or motive for a policy, but merely requires 'policies or practices that are facially neutral but have a disproportionately adverse impact on protected classes' (Barocas and Selbst 2016, 694). Thus, while the implications of a given practice for individuals or groups facing discrimination are the same under both doctrines, the hurdles for indicting someone of discrimination are substantially higher under the doctrine of disparate treatment, as it is notoriously difficult to assess intention to discriminate, let alone to prove it in court.

Please note that in both cases we are witnessing disproportionately adverse impact on protected classes, i.e. harm for those arguably already more vulnerable. What differs is the assessment of moral blameworthiness and corresponding legal liability of the employer. The core disagreement is whether employers are only liable when this harm was intended or whether it suffices that the harm occurs. Now, compare this with our four cases of deception: shall we conceive the developers of AI technologies morally responsible and thus blameworthy for the harm caused if and only if they intended to deceive, or can they be blamed for the harm caused even without necessarily having intended it?

To my mind, adopting a stance on deception which drops intent and uses the analogy of disparate impact, has the advantage of being more sensitive to harm experienced by those who are exposed to deception. It lowers the barrier to attribute responsibility to AI developers for the potential harm they may cause, just as much as it lowers the barrier for those facing discrimination in the work environment to sue their employers. While this is a conceptual choice, it is by no means unfounded: instead, it is based upon the ethical premise that power demands responsibility and that those who have more power accordingly have a higher responsibility for the consequences of their actions and non-actions, i.e. their negligence.

By adopting a broader notion of deception, which drops the requirement for intent, I do not aim to suggest that there is no difference at all between a communicative situation in which a sender of information is intending to deceive and one where a recipient is deceived without this being intended by the sender. Indeed, the intention to deceive still plays a crucial role in *assessing the degree of blameworthiness* of a (communicative) action. But it may not be relevant in assessing the practical, ethical and political consequences of being deceived.

Those unconvinced so far argue may still argue that dropping the requirement of intent muddies conceptual waters between being deceived and merely making a mistake in judgment, thereby arguing that without intention to deceive, it is the recipient of the information and she only that is to blame for being wrong about the ontological status of her AI counterpart (Deception 1), its capacities (Deception 2), for falsely believing a deepfake (Deception 3) or for confusing generative AI with a search engine (Deception 4). To my mind, however, the notion of *mistake* is not only arguably itself too broad to capture the phenomena described above: the characteristics of AI systems and the discourse surrounding them afford and thereby at least partially cause such 'mistakes'. Thus, behind what appears to be a conceptual debate about the definition of deception lurks indeed an ethical debate on how to distribute responsibility in case of doubt: to what degree are the providers of potentially misleading information responsible for the harms they cause and to what degree are the recipients of this information to be blamed for being deceived? In the following, I spell out this distribution of responsibilities in more detail, but here I want to reiterate that the conceptual choices we are making are in themselves ethical. My basic premise is that with more power comes more responsibility – and this ethical commitment has implications for my conceptual choice of adopting a wide notion of deception: a notion which drops the requirement of intent as *necessary* for

deception, while it does acknowledge that intent still matters in evaluating the *degree* of moral blameworthiness of those causing deception.

7. Conclusion: Trustworthy Generative AI & Distributed Responsibilities

If one accepts my arguments, what implications are to be drawn from these analyses on Generative AI, deception, trust and trustworthiness? Given the quadruple danger of deception posed by Generative AI and the detrimental effects of deception on trust and trustworthiness, the following argument can be made.

Trust in AI is justified if and only if, AI is trustworthy as unwarranted trust in AI can have severe epistemological and ethical consequences leading to individual, collective and societal harm (Smuha 2021). It follows that if we wish to develop and deploy Generative AI, we should aim to create Generative AI that is as trustworthy as possible. Deceptive AI reduces the trustworthiness of AI and should thus be avoided. As a consequence, there is a duty to combat the dangers of quadruple deception of Generative AI. The concrete implications of this overall duty differ depending on the type of deception and the different roles humans may have in the design, development and deployment of Generative AI, as well as the discourse around and governance of these technologies.[16] Overall, the more power actors have, the higher is their responsibility for the consequences of their actions or lack of actions.

7.1. Implications for Designing, Developing and Deploying Generative AI

Those who design and develop systems of Generative AI are the first actors coming to one's mind when thinking about possibilities and thus responsibilities to influence these technologies so that they exhibit characteristics of trustworthiness. And indeed, through design decisions, a number of potentially deceptive features of AI systems can be mitigated. One of the most central requirements concerns the avoidance of anthropomorphic design features, which can lead to various types of deception ranging from Deception 1 over Deception 2 to Deception 3. Thus, by avoiding anthropomorphism, such deceptive threats can be mitigated or at least diminished.

To further decrease *Deception 2*, it is essential that the capacities and underlying mechanisms of software, but also its limitations are properly described and communicated. Put more succinctly: the clearer it is that AI technologies are basically just advanced statistics, the less likely it is that users will ascribe consciousness or empathy to them. Knowledge about technologies and their limitations thus reduces deception and enables a more nuanced assessment of trustworthiness and a more justified way of placing – or withholding – trust in such technologies. Such knowledge can then also mitigate *Deception 4* as understanding the functionalities of software can enable one to better assess and decide how to use and how not to use different types of tools.

For the developers and providers of Generative AI, this implies a duty to adequately and truthfully communicate the basic functionalities, underlying mechanisms, as well as the limitations of their technologies. For the users, this implies a duty to use these technologies in responsible ways when using them for their various purposes – ranging from information retrieval to content generation: not only should they avoid being deceived by technology through inquiring the functionalities and limitations of the technologies they use, they are also obliged not to use such technologies in deceptive ways themselves, e.g. by creating deepfakes or cheating in exams.

And while *Deception 3* resides mostly in the hands of the users, tech companies can support the mitigation of epistemic, ethical, and societal harm here through various technological or legal means. Technological mechanisms to fight the creation or mass distribution of fake news, or at least make fake content recognizable include digital watermarks but also various tools to detect deepfakes. Thus, companies profiting from Generative AI have a moral duty to also develop and invest in technological means that can mitigate potential harm resulting from the misuse and abuse of these technologies. Moreover, developers and providers of Generative AI may influence and

reduce malicious use scenarios through licensing or usage agreements (Helberger and Diakopoulos 2023). Clearly, both the technological and regulatory mechanisms have their limitations, either because of the ensuing technological arms race or because malicious users may not be deterred by licensing agreements. Despite these and other shortcomings, these different technological and regulatory mechanisms and their effects on deception and trust should nonetheless be investigated and supported if effective.[17]

7.2. Implications for Analyzing and Discussing Generative AI

It is, however, not only the design, development, and deployment of the technologies, which can lead to deception, but also the discourse surrounding these technologies. Clearly, at times those developing or providing the technology are themselves heavily engaged in stirring the debate around their technologies and thus responsible for the deception caused by their misleading claims. News articles reporting on tech company employees who claim that their company's chatbots exhibit understanding or even consciousness,[18] as well as discussions around the various moratoria[19] which have been proposed in May 2023 serve as case studies for such deceptive debates and their detrimental effects. It is not always possible to discern whether claimants truly believe their statements or whether claiming AI agency or urgency are merely strategies to shirk responsibility and direct attention to fictitious long-term scenarios and away from harms of AI-based technologies that are already happening. As such, false and misleading claims need to be countered as they not only lead to a deceived public, but possibly also to wrong priorities regarding the governance of AI in politics – notoriously culminating in claims regarding the impossibility of governing AI by exactly those actors who are responsible for doing so.[20]

Academia clearly has a role in countering these narratives and many scholars stand up to debunk false claims about AI consciousness, agency and responsibility. However, academic debates are not entirely innocent in adding to the deception regarding AI themselves or are at least sometimes not helping in clarifying essential characteristics of artificial intelligence – even if the intention may be otherwise. It is, in particular, the all too easy attribution of agency to technological artefacts in general and AI in particular that may play into the hands of those intentionally muddying the – conceptual and political – waters on AI. At the time when theories such as actor-network theory (cf. Latour 2005; Law and Hassard 1999), mediation theory (cf. Verbeek 2005) or accounts of distributed morality (Floridi and Sanders 2004) were developed; these were set out against a prevalent ignorance of the materiality of technology within philosophy and sociology of technology. These accounts and many others thus provided a much-needed, important and nuanced corrective to earlier accounts, which neglected or downplayed the interwovenness of humans and their socio-technical environment as well as the various ways of interaction and mutual shaping and co-evolving of humans and their socio-technical environment. My critical remarks regarding the all too easy attribution of agency, responsibility or rights to AI thus are not set out against these accounts but against those more recent contributions which not only lack the originality and conceptual rigor of such early investigations, but additionally ignore the current political and societal landscape in which such contributions fall on open ears in the tech industry as they seems to justify a delegation of responsibility for tech design and its consequences away from the those building and profiting from them. Clearly, a conceptual analysis does not become false merely by being prone to abuse. However, as has been argued by feminist philosophers for a long time, we are accountable for the *agential cuts* we make (Barad 2007) and responsible for the implications of our conceptual choices (cf. Haslanger 2014).

7.3. Implications for Governing Generative AI

Finally, avoiding and countering deception regarding AI is also a responsibility for politics. First, politicians need to ensure that they are not being manipulated by misleading claims about AI or

even further, such deceptive discourse themselves through their own statements. Second, politicians need to set the parameters to ensure that deception through AI technologies is mitigated by law and policy. This first entails regulation which counters deceptive possibilities of AI, such as the labeling requirements in the AI Act. Secondly, it also requires funding and support for the development of technologies which avoid, debunk, or mitigate deceptive design. Third, it demands investment into inter- and transdisciplinary research and education to establish fruitful mutual influence and engagement between the social sciences, humanities, and legal domain on the one hand and the technological disciplines involved in AI research and design on the other. Research and education in the social sciences, humanities and legal studies needs to entail technological and mathematical competencies, while education and research in the technological domains needs to be enriched by philosophical, social, and legal insights to equip students and researchers with tools to better assess and address the societal implications of their (future) work.

7.4. Implications for Using Generative AI

Finally, it is also up to all of us as users of technologies employing Generative AI not to let ourselves be deceived by the shiny promises of Generative AI and to place – or withhold – our trust wisely.

Notes

1. Within two months, ChatGPT was used by 100 million users. For comparison: TikTok needed nine months, Instagram two years to reach the same threshold. In early 2025, ChatGPT had over 400 million weekly active users processing over one billion queries per dayCf. https://www.demandsage.com/chatgpt-statistics/(Last accessed 1 April 2025) https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app (Last accessed 1 April 2025)
2. Sometimes simpler forms of statistics are used for such software, which is still labeled as artificial intelligence. Especially for proprietary software, it is often difficult to figure out which methods of data analysis are employed. For the purpose of this article, this difference regarding the type of statistical analysis does not matter. It does matter, however, once we discuss the problems around transparency and accountability as machine learning and deep learning pose specific challenges in that regard due to the so-called *black-box problem* of AI (cf. Pasquale 2015).
3. In a previous publication, I have already outlined the danger of triple deception (Simon 2023). I am expanding this notion here by adding a fourth type of deception related to epistemic, ethical and political challenges related to the usage of Generative AI in search as well as the embedding of Generative AI infor other informational tools. For an interesting take on trust and deception regarding artificial agents that precedes the developments of Generative AI cf. Grodzinski et al. (2015).
4. For an overview on ethical issues related to using AI in psychotherapy cf Fiske, Henningsen, and Buyx (2019). The issue of deception, which is not specifically addressed in this overview has been addressed earlier in regards to robots care by Sharkey and Sharkey (2011).
5. Confer for instance https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/ (Last Accessed 1 April 2025).
6. https://correctiv.org/faktencheck/2023/03/28/mode-papst-franziskus-nein-dieses-foto-mit-weissem-daunenmantel-ist-nicht-echt/ (Last Accessed 1 April 2025).
7. https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/. and https://time.com/6565446/biden-deepfake-audio/ (Last Accessed 1 April 2025).
8. Cf. for instance, https://www.wired.com/story/the-race-to-build-a-chatgpt-powered-search-engine/ (Last Accessed 1 April 2025) or https://medium.com/web3-use-case/the-future-of-search-engines-566bbf647c62 (Last Accessed 1 April 2025).
9. It should be noted that ChatGPT and other services are indeed improving their search function by allowing users to find references and links to online sources. If clearly labeled, this will indeed mitigate the risks outlined here to some extent. However, the difficulty to distinguish which components of the output of Generative AI are *generated* and which exist elsewhere will likely remain. Accordingly, this epistemic disentanglement is likely to remain a challenge for both technology design and epistemic practices in the future.
10. For an overview of the contemporary debates on the philosophy of trust, cf. Simon (2020a).
11. To exemplify the negative effects of mis- and disinformation on trust, one may recall online communication during the Covid pandemic, where false beliefs regarding health measures have led to practical, ethical and political harms beyond mere epistemic concerns.

12. It has been argued that using chatbots in therapeutic settings may have certain merits, e.g. by increasing access and availability of basic (pre-)forms of therapy or because it may be easier for some clients to openly report their problems when interacting with a software in contrast to another human. An in-depth assessment of the pros and cons of this use of chatbots is beyond the scope of this article. Instead, I want to outline here the specific problem of deception regarding the capacities of chatbots in psychotherapeutic and similar contexts and the ethical implications to be drawn from this.
13. Cf. the following examples in which deepfakes have been used to influence public opinion and even elections in Slovakia and the US. https://www.biometricupdate.com/202401/deepfake-voice-attacks-are-here-to-put-detection-to-the-real-world-test. and https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/ (Last Accessed 1 April 2025).
14. Confer for instance, https://medium.com/coinmonks/chatgpt-could-this-be-the-future-of-search-6c8fddde4e48 (Last Accessed 1 April 2025).
15. Microsoft 365 Copilot, for instance, embeds AI into various tools for information processing and retrieval, such as Word, Excel, PowerPoint, Outlook and Teams https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/ (Last Accessed 1 April 2025).
16. Trustworthiness has been proposed as a central goal for the development of artificial intelligence before as indicated by numerous policy and academic publications (e.g. HLEG AI & surrounding literature, incl. ref to article by us). Given the brevity of this contribution, I do not assess the merits and shortcomings of this usage of the notion of trustworthiness as a normative goal for AI development or the policy initiatives here (but see G. Rieder, Simon, and Wong 2021). Instead, I focus here only on the threat of quadruple deception as it raises specific concerns not only for the design, development and deployment of trustworthy AI itself, but also for academic, public and political debates and the governance of such systems as well. These requirements and the corresponding duties for the different actors involved will be sketched in the following.
17. It can be and has been argued that licensing agreements prohibiting certain use cases can also be misused by providers to shirk responsibilities by simply prohibiting certain use cases, in particular use cases which may place their products into a high-risk class according to the AI Act. I agree that this indeed is a danger. However, all I wanted to stress here is that licensing agreements are one instrument amongst others which may have a certain if limited impact to reduce deceptive use cases.
18. Confer for instance, https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/ (Last Accessed 1 April 2025).
19. Confer for instance https://futureoflife.org/open-letter/pause-giant-ai-experiments/ or https://www.safe.ai/statement-on-ai-risk (Last Accessed 1 April 2025).
20. Cf. https://www.bbc.com/news/uk-67225158 (Last Accessed 1 April 2025).

Acknowledgements

This research was sponsored by the Volkswagen Foundation under the project "Informing Regulatory Reasoning on Algorithmic Systems in Societal Communication with STEAM – The Socio-Technical Ecosystem Architecture Method" (Az.: 9B331, 9B349).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Judith Simon is Full Professor for Ethics in Information Technologies at the Universität Hamburg. She is interested in ethical, epistemological and political questions arising in the context of digital technologies, in particular in regards to artificial intelligence. Judith Simon is Vice-Chair of the German Ethics Council, where she also was the spokesperson for the opinion on "Humans and Machines – Challenges of Artificial Intelligence". She has been serving on various other committees of scientific policy advice such as the Data Ethics Commission of the German Federal Government (2018-2019). She is the editor of the Routledge Handbook of Trust and Philosophy (2020) and serves on the editorial and advisory boards of the journals "Philosophy and Technology", "Big Data & Society" and "Digital Society".

References

Ananny, Mike, and Kate Crawford. 2016. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media and Society* 20 (3): 973–989. https://doi.org/10.1177/1461444816676645.

Angwin, J., Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks." *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bahrini, Aram, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. 2023. "ChatGPT: Applications, Opportunities, and Threats." *IEEE Systems and Information Engineering Design Symposium*. https://doi.org/10.1109/sieds58326.2023.10137850.

Baier, Annette. 1986. "Trust and Antitrust." *Ethics* 96 (2): 231–260. https://doi.org/10.1086/292745.

Barad, Karen. 2007. "Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning." https://doi.org/10.1515/9780822388128.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *SSRN Electronic Journal*. January. https://doi.org/10.2139/ssrn.2477899.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922.

Binns, Reuben. 2017. "Algorithmic Accountability and Public Reason." *Philosophy & Technology* 31 (4): 543–556. https://doi.org/10.1007/s13347-017-0263-5.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Von Arx Sydney, Michael S. Bernstein, et al. 2022. "On the Opportunities and Risks of Foundation Models." *arXiv.Org*. https://arxiv.org/abs/2108.07258.

Busch, Lawrence. 2016. "Looking in the Wrong (La)Place? The Promise and Perils of Becoming Big Data." *Science, Technology, & Human Values* 42 (4): 657–678. https://doi.org/10.1177/0162243916677835.

Coeckelbergh, Mark, and David J. Gunkel. 2023. "ChatGPT: Deconstructing the Debate and Moving it Forward." *AI & Society* 39 (5): 2221–2231. https://doi.org/10.1007/s00146-023-01710-4.

Cohen, Julie E. 2008. "Privacy, Visibility, Transparency, and Exposure." *The University of Chicago Law Review* 75 (1): 8. https://chicagounbound.uchicago.edu/uclrev/vol75/iss1/8.

Dennett, D. C. 1971. "Intentional Systems." *The Journal of Philosophy* 68 (4): 87–106. https://doi.org/10.2307/2025382.

Deutscher Ethikrat. 2023. "Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz." Accessed March 10, 2025. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf.

Fallis, Don. 2000. "Veritistic Social Epistemology and Information Science." *Social Epistemology* 14 (4): 305–316. https://doi.org/10.1080/02691720010008653.

Fallis, Don, and Kay Mathiesen. 2019. "Fake News is Counterfeit News." *Inquiry*: 1–20. https://doi.org/10.1080/0020174x.2019.1688179.

Fiske, Amelia, Peter Henningsen, and Alena Buyx. 2019. "Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy." *Journal of Medical Internet Research* 21 (5): e13216. https://doi.org/10.2196/13216.

Floridi, Luciano, and J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14 (3): 349–379. https://doi.org/10.1023/b:mind.0000035461.63578.9d.

Friedman, Batya, and Helen Nissenbaum. 1997. "Bias in Computer Systems." In *Human Values and the Design of Computer Technology*, edited by Batya Friedman, 21–40. Cambridge: Cambridge University Press.

Gelfert, Axel. 2018. "Fake News: A Definition." *Informal Logic* 38 (1): 84–117. https://doi.org/10.22329/il.v38i1.5068.

Goldman, Alvin I. 1999. *Knowledge in a Social World*. 2003 Online ed. Oxford University Press eBooks. https://doi.org/10.1093/0198238207.001.0001.

Goldman, Alvin I. 2008. "The Social Epistemology of Blogging." In *Information Technology and Moral Philosophy*, edited by Jeroen van den Hoven and John Weckert, 111–122. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511498725.007.

Grodzinsky, Frances S., Keith W. Miller, and Marty J. Wolf. 2015. "Developing Automated Deceptions and the Impact on Trust." *Philosophy & Technology* 28 (1): 91–105. https://doi.org/10.1007/s13347-014-0158-7.

Hacking, Ian. 1992. "Statistical Language, Statistical Truth and Statistical Reason: The Self-Authentification of a Style of Scientific Reasoning." In *Social Dimensions of Science*, edited by Ernan McMullin, 130–157. Notre Dame: University of Notre Dame Press.

Hardwig, John. 1991. "The Role of Trust in Knowledge." *The Journal of Philosophy* 88 (12): 693–708. https://doi.org/10.2307/2027007.

Haslanger, Sally. 2014. "Social Meaning and Philosophical Method." *Proceedings & Addresses of the American Philosophical Association* 88:16–37. http://hdl.handle.net/1721.1/97049.

Helberger, Natali, and Nicholas Diakopoulos. 2023. "ChatGPT and the AI Act." *Internet Policy Review* 12 (1). https://doi.org/10.14763/2023.1.1682.
Hendricks, Vincent F., and Mads Vestergaard. 2019. *Reality Lost. Markets of Attention, Misinformation and Manipulation*. Cham: Springer Open. https://doi.org/10.1007/978-3-030-00813-0.
Koops, Bert-Jaap. 2021. "The Concept of Function Creep." *Law, Innovation and Technology* 13 (1): 29–56. https://doi.org/10.1080/17579961.2021.1898299.
Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. 2023 Online ed. Oxford University Press eBooks. https://doi.org/10.1093/oso/9780199256044.001.0001.
Law, John, and John Hassard. 1999. *Actor Network Theory and After*. Oxford: Wiley-Blackwell.
Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2017. "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes." *Philosophy & Technology* 31 (4): 611–627. https://doi.org/10.1007/s13347-017-0279-x.
Mahon, James E. 2015. "The Definition of Lying and Deception." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, (Winter 2016 ed.). Stanford, CA: The Metaphysics Research Lab, Philosophy Department, Stanford University. https://plato.stanford.edu/archives/win2016/entries/lying-definition.
Miller, Boaz, and Isaac Record. 2013. "Justified Belief in a Digital Age: On the Epistemic Implications of Secret Internet Technologies." *Episteme* 10 (2): 117–134. https://doi.org/10.1017/epi.2013.11.
Miller, Boaz, and Isaac Record. 2016. "Responsible Epistemic Technologies: A Social-Epistemological Analysis of Autocompleted Web Search." *New Media and Society* 19 (12): 1945–1963. https://doi.org/10.1177/1461444816644805.
Minsky, Marvin L. 1968. *Semantic Information Processing*. Cambridge: MIT Press.
Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3 (2): 1–21. https://doi.org/10.1177/2053951716679679.
Nguyen, C. Thi. 2020. "Echo Chambers and Epistemic Bubbles." *Episteme* 17 (2): 141–161. https://doi.org/10.1017/epi.2018.32.
Nissenbaum, Helen. 2004. "Privacy as Contextual Integrity." *Washington Law Review* 79 (1): 119–158.
Nissenbaum, Helen. 2011. "A Contextual Approach to Privacy Online." *Daedalus* 140 (4): 32–48. https://doi.org/10.1162/DAED_a_00113.
Nyholm, Sven. 2023. "Is Academic Enhancement Possible by Means of Generative AI-Based Digital Twins?" *American Journal of Bioethics* 23 (10): 44–47. https://doi.org/10.1080/15265161.2023.2249846.
O'Neill, Onora. 2020. "Questioning Trust." In *The Routledge Handbook of Trust and Philosophy*, edited by Judith Simon, 17–27. New York: Routledge.
Pariser, Eli. 2011. *The Filter Bubble: What the Internet is Hiding from You*. London: Penguin Books. https://dl.acm.org/citation.cfm?id=2029079.
Pasquale, Frank. 2015. *The Black Box Society*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674736061.
Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?" *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.85.
Rieder, Bernhard. 2016. "Big Data and the Paradox of Diversity." *Digital Culture & Society* 2 (2): 39–54. https://doi.org/10.14361/dcs-2016-0204.
Rieder, Gernot, Judith Simon, and Pak-Hang Wong. 2021. "Mapping the Stony Road Toward Trustworthy AI: Expectations, Problems, Conundrums." In *Machines We Trust: Perspectives on Dependable AI*, edited by Marcello Pelillo and Teresa Scantamburlo, 27–40. Cambridge, MA: MIT Press. https://doi.org/10.2139/ssrn.3717451.
Scheman, Naomi. 2020. "Trust and Trustworthiness." In *The Routledge Handbook of Trust and Philosophy*, edited by Judith Simon, 28–40. New York: Routledge.
Sharkey, Amanda, and Noel Sharkey. 2011. "Children, the Elderly, and Interactive Robots." *IEEE Robotics & Automation Magazine* 18 (1): 32–38. https://doi.org/10.1109/mra.2010.940151.
Simon, Judith. 2010a. "A Socio-Epistemological Framework for Scientific Publishing." *Social Epistemology* 24 (3): 201–218. https://doi.org/10.1080/02691728.2010.498930.
Simon, Judith. 2010b. "The Entanglement of Trust and Knowledge on the Web." *Ethics and Information Technology* 12 (4): 343–355. https://doi.org/10.1007/s10676-010-9243-5.
Simon, Judith. 2015. "Distributed Epistemic Responsibility in a Hyperconnected Era." In *The Onlife Manifesto: Being Human in a Hyperconnected Era*, edited by Luciano Floridi, 145–159. Cham: Springer. https://doi.org/10.1007/978-3-319-04093-6_17.
Simon, Judith, ed. 2020a. *The Routledge Handbook of Trust and Philosophy*. New York: Routledge. https://doi.org/10.4324/9781315542294.
Simon, Judith 2020b. "Introduction." In *The Routledge Handbook of Trust and Philosophy*, edited by Judith Simon, 1–13. New York: Routledge. https://doi.org/10.4324/9781315542294.

Simon, Judith. 2023. "ChatGPT." *Deutscher Bundestag, Ausschuss für Bildung, Forschung und Technikfolgenabschaetzung." Expertengespraech zum Thema ChatGPT, Ausschussdrucksache* 20 (18): 108b. Accessed March 10, 2025. https://www.bundestag.de/resource/blob/944448/004ca2f7a9fcf586a07113c6ba72b689/20-18-108b-Simon-data.pdf.

Simon, Judith, Pak-Hang Wong, and Gernot Rieder. 2020. "Algorithmic Bias and the Value Sensitive Design Approach." *Internet Policy Review* 9 (4): 1–16. https://doi.org/10.14763/2020.4.1534.

Simpson, John, and Edmund Weiner, eds. 1989. *The Oxford English Dictionary*. Oxford: Clarendon Press.

Smuha, Nathalie A. 2021. "Beyond the Individual: Governing AI's Societal Harm." *Internet Policy Review* 10 (3). https://doi.org/10.14763/2021.3.1574.

Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind* LIX (236): 433–460. https://doi.org/10.1093/mind/lix.236.433.

Verbeek, Peter-Paul 2005. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. University Park: Penn State University Press.

Weizenbaum, Joseph. 1966. "ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM* 9 (1): 36–45. https://doi.org/10.1145/365153.365168.

Weizenbaum, Joseph. (1977) 1994. *Die Macht der Computer und die Ohnmacht der Vernunft*. Translated by Udo Rennert. Frankfurt: Suhrkamp.

Weizenbaum, Joseph. 2001. *Computermacht und Gesellschaft: Freie Reden*. Frankfurt: Suhrkamp.

World Economic Forum. 2024. The Global Risks Report 2024: Insight Report. https://www.weforum.org/publications/global-risks-report-2024/.