AI-Testimony, Conversational AIs, and Our Anthropocentric Theory of Testimony

Ori Freiman[1] (Digital Society Lab, McMaster University, Canada)

The ability to interact in a natural language completely changes devices' interfaces and potential applications of speaking technologies. Concurrently, this phenomenon challenges our mainstream theories of knowledge, such as how to analyze linguistic outputs of devices under existing anthropocentric theoretical assumptions. In section 1, I present the topic of machines that speak, connecting between Descartes and Generative AI. In section 2, I argue that accepted testimonial theories of knowledge and justification commonly reject the possibility that a speaking technological artifact can give testimony. In section 3, I identify three assumptions underlying the view that rejects conversational AIs – AI-based technologies that converse, as testifiers: conversational AIs (1) lack intentions; (2) cannot be normatively assessed; and (3) cannot constitute an object in trust relations; while humans can. In section 4, I propose the concept 'AI-Testimony' for analyzing outputs of conversational AIs, suggesting three conditions for technologies to deliver AI-testimony: (1) content is propositional; (2) generated and delivered with no other human directly involved; (3) the output is perceived as phenomenologically similar to that of a human. I conclude that this concept overcomes the limitations of the anthropocentric concept of testimony, opening future directions of research without associating conversational AIs human-like agency.

1. Introduction and Overview

The father of modern Western philosophy, Rene Descartes, thought it impossible for a machine to make meaningful verbal responses to verbal stimuli. He addressed this topic in his famous *Discourse on the Method:*

> We can certainly conceive of a machine so constructed that it utters words, and even utters words which correspond to bodily actions causing a change in its

[1] freimano@mcmaster.ca
https://orcid.org/0000-0002-6750-9130

> organs (e.g., if you touch it in one spot it asks what you want of it, if you touch it in another spot it cries out that you are hurting it, and so on). But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do (Descartes 1985 [1637]: 140)[2].

Of course, history - through the development of technology - has proven Descartes wrong. Since Alan Turing's (1950) famous paper about the imitation game, research in conversations between humans and computers was sparked (Licklider & Robert 1968; Voicebot AI 2019).

Fast forward to the second decade of this century. An example of a conversation technology is Google Duplex. It is "an AI system for accomplishing real-world tasks over the phone" (Google AI Blog 2018). This technology "is built to sound natural" (ibid), and it does:

> It sounds remarkably - maybe even eerily - human, pausing before responding to questions and using verbal tics, like 'um' and 'uh.' It says 'mm hmm' as if it's nodding in agreement. It elongates certain words as though it's buying time to think of an answer, even though its responses are instantaneously programmed by algorithms (Nieva 2018).

At the end of November 2022, OpenAI released their chatbot –'ChatGPT' (OpenAI 2022). This product popularized the Generative AI technology – a type of AI that can generate new forms of content such as audio, images, videos, and texts - based on learning from existing data using neural networks. Generative AI is considered by many as a technological breakthrough, and so does its flagship product – ChatGPT, since they can be utilized to perform a wide range of tasks, from writing love songs and summarizing academic texts, to suggesting creative ideas, coding, analyzing political events, and so much more.

The surprising ease of use and wide range of possibilities led to the rapid adoption of technology, hitting one million users in five days (Mollman 2022; Hurst 2022). During the

[2] This passage was brought to my attention by Nickel (2013a), who credits his attention to Salinga & Wuttig (2011).

beginning of this hype cycle, ChatGPT was constantly in the media spotlight, drawing much attention from the public, industry, scholars, and policymakers. Developments have even led to public and expert discussions about the threats of Artificial General Intelligence (Future of Life Institute 2023; Center for AI Safety 2023; cf. Bender 2023; DAIR 2023).

Conversational AIs are technologies that interact with humans using natural language. These technologies improve at a sensational pace. The technological abilities to generate textual output, receive audible inputs from human speech, and communicate the generated content in audible and written formats constantly advances. When the linguistic fluency of Generative AI is combined with speech technologies, conversations with technological devices become undistinguishable, sound-wise, and arguably soon also content-wise, from conversations with a human. The ease of use - natural language, renders this technology a potential interface for other technologies. Conversational AIs and LLMs already pose new societal challenges (Ruane, Birhane, & Ventresque 2019; Bender et al. 2021; Mökander et al. 2023). One example is the changing landscape of customer service - making access to human service representatives a privilege, limited exclusively to paid customers or those who have subscribed (Geslevich Packin & Freiman 2023).

The ability to interact in a natural language completely changes devices' interfaces and our societies by posing a promise to further revolutionize the collection and consumption of information. At the same time, speaking with conversational AIs challenges our mainstream theories of knowledge. This paper deals with one such challenge within the field of social epistemology, which is raised by testimonial theories of knowledge and justification: how to characterize the commonly held view of testimony.

2. Can a Conversational AI Deliver Testimony?

Can a conversational AI deliver testimony? The concept of testimony plays a crucial role in the sociology and the history of sciences. Shapin's (1994) study of knowledge production in seventeenth-century England shows that the scientific culture of that time was built upon the word of a gentleman. Social status was a crucial factor when considering whom to trust. Unlike the common merchants or laborers, gentlemen were not affected by economic pressure that could compromise their ability to tell the truth. Then, so as now, the testimony that we accept is also based on social factors.

Shapin argues that "knowledge have a moral character" (Shapin 1994: xxv). According to his view, instruments lack a moral character that gentlemen can have and therefore cannot, in principle, give testimony. Shapin's work highly influenced the development of the theory of testimony[3].

The common view of testimony holds, generally, that only persons participate in the act of testimony. The view that only persons participate in the act of testimony is advocated by leading sociologists (e.g. Bloor 1999; Collins 2010; Collins & Kusch 1998), and also rooted in mainstream philosophical accounts of testimony (Coady 1992: 268; Lackey 2008: 189).

Nevertheless, we do acquire beliefs from technological artifacts that speak. When I hear the voice of my car's navigation system telling me that "road number 31 is closed", I form a belief that "road number 31 is closed". Even further: after talking with ChatGPT about the impact of immigration on society, I changed my views on affirmative action (roughly, a policy that aims to increase the participation of groups that have faced historical disadvantage or discrimination). Despite the ability to acquire beliefs and change one's views, a common assumption differentiates algorithms from humans as potential testifiers: algorithms, similar to instruments, lack a moral character. How, then, would we analyze a simple case of a human conversing with a virtual assistant, chatbot, robot or any other AI system built to have a conversation?

In the context of technology, Miller & Record (2013: 121fn3) note that "the question of whether and on what conditions information from computers and other instruments constitutes testimony has been largely overlooked". On the one hand, mainstream assumptions that are accepted in social epistemology maintain a sharp and normative

[3] Later on, the work of the Strong Programme in the Sociology of Knowledge was influenced from Shapin's work. They advocated a theory of knowledge that recognizes testimony as a way for epistemic communities to generate knowledge. This is a result of the ability of testimony to transforms opinions into knowledge – and establish social agreement (Kusch 2002; for further sociological context of the concept of testimony, see Neges 2018 and Freiman & Miller 2021). In fact, it might be that philosophical interest in the concept of testimony was the result of influence from sociology and the history of sciences (Clément 2010), from works such as Latour and Woolgar (1986), Latour (1988), Shapin & Schaffer (1985), and Shapin (1994).

distinction between humans and technological artifacts. This distinction is expressed in the common view that rejects the possibility of testimony by a conversational AI. On the other hand, the practice of acquiring knowledge from a conversational AI gradually becomes more and more common, leaving a lacuna in testimonial theories of knowledge.

3. Testimonial Theories are Anthropocentric

Two concepts were considered by Freiman (2023) for the analysis of beliefs acquired from conversational AIs: 'testimony-based beliefs' and 'instrument-based beliefs'. He argued that both concepts fail to correctly capture acquiring beliefs from conversational AIs: While the concept of instrument-based belief acknowledges the non-human agency of the source of the belief, the content is discussed in terms of signs and indicators and not natural language. At the same time, the concept of testimony-based belief *does* refer to natural language propositions, however, the underlying assumption is that the agency of the testifier is a human[4]. He calls this view the 'anthropocentric view of testimony' and identifies three components of this underlying assumption:

> The view presupposes that only persons can participate in the act of testimony because only humans, in principle, can be qualified as a testifier. Underlying this view are commonly held assumptions in mainstream social epistemology that a testifier (a) must have intentions to deliver the testimony; (b) be subject to normative assessment; and (c) constitute a putative object in trust relations. (Freiman 2023)

The dominant view of testimonial theories of knowledge and justification entails anthropocentric assumptions about the testifier. While scholars sporadically defend different assumptions regarding the demand that only persons participate in the act of testimony, in that paper Freiman (ibid) bundles the different anthropocentric assumptions together. In the rest of this section, I further focus on the three assumptions mentioned

[4] Instead of the concepts of testimony-based beliefs and instrument-based beliefs, Freiman (2023) suggested a third: technology-based beliefs. It is a hybrid concept, acknowledging the non-human agency of the originator of the belief that also focuses on natural language propositions that forms the subject's beliefs.

above. My purpose is to elaborate on them, and point to the differences between persons and technological artifacts.

3.1. Intentions

The first assumption is that a speaker cannot give testimony unless they intend to. Following this assumption is that it is categorically impossible for conversational AIs to deliver testimony - since they lack intentions to deliver it.

When it comes to beliefs, intentions are regarded as either reduced to some kind of desire, or some vague capacity to know what one is doing. It varies between what G. E. M. Anscombe (1963: 50) refers to as practical knowledge without observation, to what is referred to by Paul Grice (1971: 268) as "licensed wishful thinking" that does not only depend on one's will, but also on premises about one's own abilities (both cited in: Setiya 2018: §5).

Considerations of intentions in the theory of testimony are examined by Jennifer Lackey (2008) that points to the dual nature of testimony: "On the one hand, testimony is often thought of as an *intentional act on the part of the speaker* and, on the other hand, testimony is often thought of as simply a *source of belief or knowledge for the hearer*." (2008: 3, italics in origin). After considering different views[5], she advocates for the Disjunctive View of the Nature of Testimony, that captures these two independent aspects:

> S testifies that *p* by making an act of communication *a* if and only if (in part) in virtue of *a*'s communicable content, (1) S reasonably intends to convey the information that *p* or (2) *a* is reasonably taken as conveying the information that *p*. (Lackey 2008: 35-36)

A generous interpretation of the Disjunctive View of the Nature of Testimony might focus on condition (2), arguing that the communication from the conversational AI conveyed information that *p*. However, Lackey (2006, 2008) also demarcates devices as non-capable,

[5] Among the considered views are those of Graham (1997: 227) and Coady (1992: 42), both requiring intention to deliver testimony.

in principle, to deliver testimony. Lackey's threshold for participating in the act of testimony demands qualities that are unique to persons[6]: "… testimonial beliefs are acquired from *persons*. Persons, unlike other sources of belief, have all sorts of different intentions, desires, goals, motives, and so on" (2006: 176; 2008: 189, italics in origin). Under the assumption that conversational AIs are incapable of having intentions, they *ipso facto* lack the necessary requirement for delivering testimony.

Other exemplar proponent of this view is Fricker (2015), referring to pre-recorded sound of a human voice from a non-human agent as a "fake testimony":

> This contrast of natural and epistemic kind between natural versus agential meaning is muddied by the existence of what I think of as fake testimony: announcements at railway stations of train times, or automated messages one receives on telephone connections, that sound like a live human voice making statements, but are no such thing (2015: 179).

In this case these are recorded announcements, and not AI generated propositions and voices. However, Fricker uses this case to distinguish between *natural* and *agential* speakers - natural speaker who can deliver testimony, and an agential speaker whom she thinks of as giving "fake testimony".

To solve the problem of belief formation due to the speech by the artificial agent, Fricker does not undermine the knowledge gained by the hearer: "Finding out that the utterances are produced by an automated artificial speaking mechanism does not, in this case, undermine the basis she has for believing what she hears" (2015: 201). On the one hand, Fricker rejects these cases as testimonies; on the other hand, she acknowledges belief formation due to artificial speech by technological devices. This is what Freiman (2023) calls technology-based beliefs.

To conclude, I take the "paradigm case of testimony - the intentional transfer of a belief from one agent to another" (Pritchard 2004) to represent the common view about intentions within testimonial theories of knowledge and justification. Under the assumption

[6] Lackey uses the term 'person' in a broad sense, that includes non-human animals (2008, fn 13).

that technologies do not have intentions, conversational AIs cannot give testimony - in principle. Therefore, the concept of testimony is not suitable for the analysis of acquiring knowledge from technologies.

3.2. Normative Assessment

The second assumption underlying the rejection of testimony by conversational AIs comes from a categorical demand that a testifier is assessable from a normative perspective. In addition to Shapin's demand that testimony can only be given by a testifier who has a moral character (§2), I take the demand for normative assessment to mean that the testifier can be assessed as a rational and responsible epistemic agent (Goldberg 2012: 188).

A commonly held view is that a person who tells something is normatively responsible for the truth of what they tell (Fricker 2002: 379). A similar view is further defended by Sanford Goldberg (2012): in his view, only testimony that originate from humans, i.e. testimony-based beliefs, testimony that originates from an epistemic agent who is "susceptible to full-blooded normative assessment" (191) and "sophisticated enough to satisfy the conditions on being appropriately assessed in terms of rationality and responsibility" (194).

Instead, normatively assessment is often reduced to the individuals and groups behind the technologies. For example, "if a given computer yields information that turns out to be false, we will blame the programmer, or our use of the program, or ..., but not the computer itself" Goldberg (2012: 194). Indeed - in the era of AI, and especially with generative AI, the accountability and responsibility for the truth and the consequences of the output are distributed among many hands (e.g. Slota et al. 2021). Organizations developing AI-based products are expected to be responsible and accountability for the products, and the organizations and individuals are to be normatively assessed (Freiman & Geslevich Packin 2022).

The view expressed by Goldberg and Fricker can be considered to be the common view of testimony regarding normative assessment. Following this view, then, it is not possible to normatively assess conversational AIs, since they are not rational or responsible epistemic agents. Therefore, the conversational AI cannot give testimony.

3.3. An Object of Trust

The concept of trust closely relates to testimonial theories mostly because testimonial accounts of knowledge require some kind of trust relations between the speaker and the hearer (e.g. Faulkner 2011; Gelfert 2014: §8, 2018: §5). A third reason for rejecting the possibility that conversational AIs can give testimony comes from a categorical demand that the testifier be an object in trust relations.

Mainstream views within the epistemology of trust point to the concept of 'reliance' to correctly describe expectations from machines to perform well (Baier 1986; Nickel 2013b). Similar to the epistemology of testimony, in the traditional paradigm of trust within analytic philosophy, genuine trust is based on human quality (Miller & Freiman 2021; Freiman 2022).

This view of trust mostly rests on Annette Baier's (1986) phrasing of the idea that "trusting can be betrayed, or at least let down, and not just disappointed" (235), and that people, but not artifacts, can betray. Karen Jones (1996) developed Baier's account of trust, arguing that it is not possible to genuinely trust technologies: "trusting is not an attitude that we can adopt toward machinery" (14). According to this view, technological artifacts cannot be objects of trust, not because they are inherently untrustworthy, but because trust relations cannot be formed with non-persons, in principle.

Similar to the reduction of normative assessment to the individuals and organizations, when philosophers discuss trust in technologies, they usually reduce the discussion to trusting the humans behind the technologies. For example, when a person trusts a bridge not to collapse, she actually trusts the people who built the bridge and the people who are responsible for its maintenance (Origgi 2008).

In the context of AI, Freiman (2022) points to philosophers and ethicists of AI that are aware of the anthropocentricity in light of the commonly used concept of 'Trustworthy AI' in policy documents and research. His analysis of trust elucidates why the concept of 'Trustworthy AI' has been labelled as a "conceptual nonsense" (Metzinger 2019), "conceptual misunderstanding" (Hatherley 2020), and "a misnomer" (Braun, Bleher, & Hummel 2021) among those who philosophize, arguing that the social-epistemic concept of trust entails anthropocentric assumptions.

4. AI-Testimony from a Conversational AI

Acknowledging the anthropocentric nature of the social-epistemic concept of testimony still leaves the challenge of how to analyze the technology-based belief a person acquires from a conversational AI.

In prior work on the concept of testimony in the context of technology, Freiman & Miller (2021) focused on the differences in assertion between humans and machines. They argue that machine assertion currently cannot understand puns, cynicism, or humor; or skillfully answer an out-of-context question. Additionally, they illustrate that, unlike humans, computers are not (yet) sensitive to small errors that humans will immediately notice.

Freiman & Miller (2021) build upon Nickel (2013a) and Wheeler (2020) to introduce the notion of ‘quasi-testimony’, but do not further develop that notion. Taking a cue from their work, and to continue research on human perceiving natural language outputs by technology and specifically conversational AIs, I suggest the concept of ‘AI-testimony’. With this concept I emphasize the non-human agency of the source of the output that is delivered.

In the rest of this section, I further develop the concept of AI-testimony. First, I show how this concept can be complementary to, and compatible, with anthropocentric theories of testimony, discussing intentions and reduction of trust and normative assessment to humans. I then I suggest three conditions for a conversational AI to deliver AI-testimony.

4.1. Compatibility with Anthropocentric Theories of Testimony

In section 3, I characterized testimonial theories as anthropocentric, by maintaining that mainstream social epistemology assumes a testifier must have intentions to deliver the testimony, be subjected to normative assessment, and be a valid object in trust relations. Despite the non-human agency of AI-testifiers, these assumptions must be dealt with.

I agree that technologies lack the ability to have intentions. However, I suggest that an analysis of knowledge from conversational AI need not ascribe the source of knowledge to an agent with intentions. Within the theory of testimony, arguments exist that intentions are not necessary for delivering a testimony. Coady (1992: 51) refers to such examples in his "documentary testimonies" - reading a personal diary written by someone else, and Lackey (2008: 18) gives the example of Sylvia Plath's posthumously published diaries. In these cases, the author had no direct intention that anyone would read the text they authored.

There is no reason why we cannot discuss the production of propositions by algorithms as a mere technical process, without associating it with human intentions. Instead, we can still discuss "presumption of human involvement somewhere in the process" (Gelfert 2014: 27) of the production of the propositions.

When people, instruments, algorithms, and other things affect the production of an epistemic outcome, the problem of assigning individual epistemic responsibility arises. This is the epistemic problem of 'many hands'. Helen Nissenbaum characterizes the problem as follows:

> Where a mishap is the work of "many hands," it may not be obvious who is to blame because frequently its most salient and immediate causal antecedents do not converge with its locus of decision-making. The conditions for blame, therefore, are not satisfied in a way normally satisfied when a single individual is held blameworthy for a harm (1996: 29).

The framework of AI-testimony cashes out the responsibility for the truth value of the propositions communicated, in terms of a 'many-hands' account. In an ever-more complex socio-technical systems, it is difficult to locate and attribute responsibility (Simon 2015: 145). Since it is categorically impossible to hold an artifact such as a conversational AI epistemically responsible (as we can hold a person) then the problem is twofold: (a) not being able to assign epistemic responsibility to a conversational AI, and (b) correctly locating and attributing epistemic responsibility among those who take part in producing the epistemic outcome.

Accounts exist that reduce questions of trust, reliability and responsibility in technologies to the humans behind them. Since trust involves some kind of confidence that a subject will be epistemically responsible in providing true statements, both the concepts of trust and responsibility demand human agency.

In the epistemology of trust, it is common to reduce responsibility, whether for the maintenance of the bridge or for the truth outcome of a software, to the humans and institutions behind. This reductionism had received various labels in the epistemology of trust (see Freiman 2022 and references within). More broadly, reducing trust and responsibility to the 'humans behind the machine' is prominently argued and defended by Joseph Pitt, who coined the sentence "technology is *humanity* at work" (2010: 445, emphasis in original; see also: Pitt 1983). Mark Coeckelbergh (2012) characterizes this approach as "direct trust in artefacts is indirect trust in the humans related to the technology". Philip J. Nickel's (2013b) *entitlement account of trust in technological artifacts and socio-technological systems* and Sanford Goldberg's (2016) *epistemically engineered environments* are examples for dealing with complex socio-technical systems without assigning non-humans the epistemic responsibilities of humans.

The concept of AI-testimony does not pretend to solve the problem of assigning responsibility in the 'many hands' problem; it leaves the dynamics of how exactly responsibility and trust are distributed for future work. In contrast to the concept of testimony, AI-testimony does not require the conversational AI to be able to have intentions to deliver the proposition, be normatively assessed, or be an object in trust relations.

In my view, these are human characteristics that cannot be applied directly to artifacts. However, analysis of intentions, normative assessment, and being an object in trust relations can, and should, be reduced to the humans and institutions behind the technologies. In this respect, the concept of AI-testimony is complementary to the existing concept of testimony, allowing a broader spectrum of analyses for linguistic outputs – including non-humans.

4.2. Three Conditions for a Conversational AI to Deliver AI-Testimony

In the rest of this section, I propose three main conditions for a conversational AI to deliver AI-testimony: (a) The output is delivered in propositions; (b) the propositions were generated by AI and delivered algorithmically, with no direct human involvement; and (c) there is a phenomenological similarity between perceiving the AI-testimony from a conversational AI and a testimony from a human.

a. Propositional Content is Delivered or Inferred

The first condition is that the output is delivered in propositions. While conversational AIs meet this demand easily, that is - they speak in natural language, not all forms of communication are done with propositions and can be considered as AI-testimonies. For example, think of a humanoid robot performing a hand gesture such as the ‘okay’ or ‘stop’ hand signs that humans sometimes do. Therefore, a weaker demand allows for inferring propositions.

This weaker demand works as long as the output entails a shared language (under existing norms of its epistemic community). Taking a cue from the theory of testimony – in the right context, propositions can be inferred also from signs, symbols, pictures, and images. For example, think about a driver who hears a siren, which usually leads to inferring from the non-verbal audible tones the proposition "This is an emergency” and the injunction “Clear the way!” (or an equivalent). Other kind of testimonies, such as reading maps and road signs can amount to testimony as well (“documentary testimonies”, see Coady 1992: 51; see also “institutional testimonies” in 1992: 50, 87).

The output of the AI-testifier must carry a proposition, either in speaking with natural language, or, under the weaker demand - with an agreed non-verbal cue that a proposition can be inferred from.

b. Propositions Were Generated by AI and Delivered Algorithmically

A second condition that must be met for analyzing the output that *p* as an AI-testimony requires the content of the proposition to be generated by AI, and be delivered algorithmically – without a direct human command. This is to contrast cases of AI-testimony from two cases: pre-made propositions uttered by artificial speakers, and a technological device that mediates human testimony.

Beginning with the first case: some artificial speakers would satisfy the first condition, but not the second. For example, when a modern defibrillator gives the vocal order "Attach the sticky pads to the patient's skin". Similarly, cases of automatic announcements at railway stations of train times (Fricker 2015) or the automatic human-recoded phone messages (Green 2006) are indeed delivered automatically, but as long as they were not generated by an AI (but pre-recorded), they cannot be considered to be AI-testimony. Contrarily, if algorithms constructed the proposition[7] and the other conditions spelled here are met, then this is an AI-testimony. While artificial speakers are technological devices that speak – sometimes pre-recorded messages, conversational AIs algorithmically generate the content of their speech.

The second class of cases that this condition comes to separate from is the mere human testimony that was mediated through devices. For instance, an exchange of text messages between two humans through their smartphones would not count as an AI-testimony. When my smartphone reads aloud a text message I received from my brother, it is not an AI-testimony but rather a testimony from my brother that is mediated by our smartphones (and the whole supporting infrastructures). The technical causal chain from my brother's smartphone to the screen and speakers of my own smartphone is extremely sophisticated and is nothing but direct. Yet, the content of the text message was thought of and sent as a direct consequence of my brother's actions.

[7] The proposition "The train from Union station will enter platform 4, in 3 minutes" can preexist as a whole, or can be assembled by an algorithm picking up information from different sources: "The train from" + X + "will enter platform" + Y + "in" + Z + "minutes".

Two examples can draw some of the spectrum of devices that their outputs can be considered as AI-testimony: Helen, a virtual assistant of the Cosibot ("COvid Stay Informed Bot") initiative[8] can tell me (voice or text) how many COVID-19 vaccine doses have been administered to date ("as of 21 June 2021, 2,414,347,324 vaccine doses have been administered"). In this case, Helen's answer is generated automatically, drawing on data from the *WHO Coronavirus (COVID-19) Dashboard.* Helen's answer was generated and delivered by algorithms, with no direct human involvement. No human thought of the specific content of the proposition or that it should be sent to me. Therefore, a virtual assistant that generates the propositions can be considered as an AI-testifier.

The second example is much more complex and gradually becomes more common: Large Language Models (LLMs). LLMs are mathematical models of the statistical distribution that recognize, predict and generate text. These models use machine learning and other AI techniques to calculate and determine the probability of text occurrence. In a nutshell, the models are trained on a vast amount of texts to learn patterns and relationships (Shanahan 2022).

There is much criticism about LLMs, touching aspects ranging from copyright, accuracy, bias, misinformation, to environment and behavior (Harrer 2023). Unlike a view that LLMs are sentient – a view that was popularized among some of the wide public and engineers by ex-Google engineer Blake Lemoine (Tiku 2022; cf. Bryson 2022), for many philosophers it is clear that LLMs are not sentient, do not have minds, their lingual outputs are not the expression of thoughts, but are rather only sophisticated statistical machines (Bender et al. 2021). LLMs are not ethical reasoners (Albrecht, Kitanidis, & Fetterman 2022), and are "intelligence-free" (Floridi 2023) in the sense that their outputs "have nothing to do with the cognitive processes present in the animal world and, above all, in the human brain and mind, to manage semantic contents successfully" (ibid).

With criticism accompanying the implementation of this technology, LLMs have demonstrated exceptional performance across many natural language processing (NLP)

[8] The Covid Stay Informed Bot initiative is based on answers from reputable sources. It is available at https://cosibot.org.

tasks. There are potential applications for LLMs in many fields - from prediction of stock price movements (Lopez-Lira & Tang 2023), to helping scientists predict protein structures (Lin et al. 2023); and for individuals with endless numbers of tasks. With the ability to generate human-like content - it can be considered a game-changer for many technologies, that will converge with conversational AIs. As this technology is adopted, propositions that will be generated with LLMs are likely to be considered as AI-testimony.

c. Phenomenological Similarity to Humans

A third condition that a machine's output must meet for us to be able to analyze its output as an AI-testimony that p - is that delivery of the output by the AI-testifier is phenomenologically similar to human delivery of a similar output . This condition is a matter of a degree.

For example, the equivalent of receiving a text message generated by a human would be receiving a text message generated by an AI. An equivalent of talking with a human telling me in a calm tone - that the goal of mindfulness is to cultivate a heightened awareness of the present moment, would be talking with a humanoid, that would tell me the same and in the same tone. Speaking face-to-face with a humanoid demonstrates much more phenomenological similarity to a human than chatting though text messages, and therefore phenomenological similarity is a matter of a degree.

Phenomenological similarity is tied with anthropomorphism: Anthropomorphizing is ascribing human-like features to a non-human. Anthropomorphic cues are interpreted as human-like and cause response, as if the technological artifact were human (Lee 2008). Anthropomorphic design can make machines simpler to use (Moon & Nass 1996), and therefore a design goal of many technologies, such as companion robots.

Recent applications of LLMs, such as OpenAI's ChatGPT and Microsoft's Bing's chatbot, present their output in an anthropomorphic manner. For example, the answer is slowly typed as if a human types it. Moreover, these chatbots use emojis that causes emphatic responses. Anthropomorphic impression, especially emotive language expressed by products, are argued to be manipulative (Véliz 2023). Interactive avatars, such as those created by D-ID, are "giving a face to conversational AIs" (D-ID 2023). These avatars look and sound real, creating an immersive human-like experience. They present high phenomenological similarity to a human being.

5. Summary and Conclusion

To sum up, for analyzing a conversational AI's output as an AI-testimony, three main requirements must be met: first, that output is delivered in propositions. Second, that output is generated by AI and delivered algorithmically, with no other human directly involved. Third, the conversational AI demonstrates a phenomenological similarity to humans.

After spelling out the conditions that a conversational AI must meet to give an AI-testimony that *p*, it is now possible to define AI-testimony:

> [AI-Testimony] A human subject *S* obtains knowledge that *p* from a conversational AI *C* by AI-testimony, if (i) *C* produces an output *O* that states a proposition *p* or that *p* can be inferred from *O*; (ii) *O* is generated by an AI and *O* is delivered in an algorithmic process in which no other human is directly involved; (iii) *S* perceives *O* from *C* in a phenomenological similar way to how *S* would perceive a testimony that *p* from another human; (iv) and *p* is true;

A concept of AI-testimony can be used to successfully analyze a human subject's acquisition of knowledge or justified belief from conversational AI without associating a human-like agency to the conversational AI, or undermine assumptions about intentions, normative assessment, being an object of trust, or any other difference between persons and technological artifacts.

Future research about AI testimony should further integrate with existing social epistemic literature, to analyze novel technologies that autonomously generate and disseminate propositions. Examples for literature include failure to believe speakers due to inappropriate prejudices (Fricker 2007), challenges of content moderation (Frost-Arnold

2019), false inferences (Alfano & Skorburg 2018), taking advantage of dishonest anthropomorphism to surveil and manipulate humans (Danaher 2020), delivering fraud messages (Waddell 2019), using voice synthesis for social engineering (Bendel 2017), the spread of misinformation (Yee 2023), and many others. Such a research direction can bypass the theoretical limitations that exist within the anthropocentric view of testimony.

As humans gradually communicate more with AIs, it becomes crucially important to maintain the distinction between humans and AIs as sources of knowledge. This paper makes two steps in this direction: it uncovers the anthropocentric assumptions in the concept of testimony - that prevents analysis of speaking technologies; and it suggest an alternative route to further research – the concept of AI testimony – that can be applied on conversational AIs.

Disclosure Statement

The author report there are no competing interests to declare.

Acknowledgement

[Will be added]

References

Albrecht, J., Kitanidis, E., and Fetterman, A. J. (2022). Despite" super-human" performance, current LLMs are unsuited for decisions about ethics and safety. arXiv preprint arXiv:2212.06295.

Alfano, M., and Skorburg, G. (2018). Extended knowledge, the recognition heuristic, and epistemic injustice. In: Carter, Clark, Kallestrup, Palermos, and Pritchard (eds.) Extended Epistemology. Oxford Scholarship Online.

Anscombe, G. E. M. (1963). *Intention.* 2nd edition. Oxford: Blackwell.

Baier, A. (1986). “Trust and antitrust”, *Ethics* 96(2): 231-260.

Bendel, O. (2017). “The synthetization of human voices”, *AI & Society* 34(1): 83-89.

Bender, E. M. (2023). Policy makers: Please don’t fall for the distractions of #AIhype. https://medium.com/@emilymenonbender/policy-makers-please-dont-fall-for-the-distractions-of-aihype-e03fa80ddbf1

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?□. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Bloor, D. (1999). "Anti-Latour", *Studies in the History and Philosophy of Science* 30(1): 81-112.

Braun, M., Bleher, H., and Hummel, P. (2021). A leap of faith: is there a formula for “Trustworthy” AI?.

Bryson, J. (2022). One Day, AI Will Seem as Human as Anyone. What Then?. Wired Jun 26, 2022. https://www.wired.com/story/lamda-sentience-psychology-ethics-policy/

Center for AI Safety. (2023). Statement on AI Risk. https://www.safe.ai/statement-on-ai-risk

Clément, F. (2010). “To trust or not to trust? Children’s social epistemology”, *Review of Philosophy and Psychology* 1(4): 531-549.

Coady, C. A. J. (1992). *Testimony: A Philosophical Study.* Oxford University Press.

Coeckelbergh, M. (2012). “Can we trust robots?”, *Ethics and Information Technology* 14(1): 53-60.

Collins, H. M. (2010). “Humans not instruments”, *Spontaneous Generations: A Journal for the History and Philosophy of Science* 4(1): 138-147.

Collins, H. M., and Kusch, M. (1998). *The Shape of Actions: What Humans and Machines Can Do*. MIT Press.

D-ID. (2023). Experience the future of conversational AI with D-ID. https://www.d-id.com

DAIR. (2023). Statement from the listed authors of Stochastic Parrots on the "AI pause" letter. https://www.dair-institute.org/blog/letter-statement-March2023/

Danaher, J. (2020). "Robot Betrayal: a guide to the ethics of robotic deception", *Ethics and Information Technology* 22(2), 117-128.

Descartes, R. (1985). "Discourse on the Method." In: *The Philosophical Works of Descartes,* vol. I. (J. Cottingham, R. Stoothoff, and D. Murdoch, Translators). Cambridge University Press.

Faulkner, P. R. (2011). *Knowledge on Trust*. Oxford University Press.

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy and Technology*, 36(1), 15. https://doi.org/10.1007/s13347-023-00621-y

Freiman, O. (2014). "Towards the Epistemology of the Internet of Things Techno-Epistemology and Ethical Considerations Through the Prism of Trust" *The International Review of Information Ethics*, 22, 6-22.

Freiman, O. (2021). The Role of Knowledge in the Formation of Trust in Technologies. PhD Dissertation, Bar-Ilan University.

Freiman, O. (2022). "Making sense of the conceptual nonsense 'trustworthy AI'", *AI and Ethics*, 1-10.

Freiman, O. (2023). "Instrument-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs: Analysis of Beliefs Acquired from a Conversational AI", *Episteme*.

Freiman, O. and Geslevich Packin, N. (2022). 'Artificial Intelligence Products Cannot be Moral Agents.' Toronto Star, 7 August. https://www.thestar.com/opinion/contributors/2022/08/07/artificial-intelligence-products-cannot-be-moral-agents-the-tech-industry-must-be-held-responsible-for-what-it-develops.html.

Freiman, O., and Miller, B. (2021). "Can artificial entities assert?", In: S. C. Goldberg (ed.), *Oxford Handbook of Assertion*, pp. 415-434. Oxford University Press.

Fricker, E. (2002). “Trusting others in the sciences: a priori or empirical warrant?”, *Studies in History and Philosophy of Science Part A* 33(2):373–383.

Fricker, E. (2015). “How to Make Invidious Distinctions Amongst Reliable Testifiers”, *Episteme* 12(2):173-202.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Frost-Arnold, K. (2019). "Epistemic Injustice and the Challenges of Online Moderation", Invited keynote lecture at Knowledge in a Digital World: Epistemic Injustice, Bias, and other Challenges in the Age of Artificial Intelligence, Canadian Society for Epistemology, Montreal, November 2019.

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

Gelfert, A. (2014). *A Critical Introduction to Testimony*. Bloomsbury Publishing.

Gelfert, A. (2018). *Testimony*. Routledge.

Geslevich Packin, N. and Freiman, O. (2023). Automation's Hidden Costs: The Case Against A Paywalled Human Touch. Forbes, May 22, 2023. https://www.forbes.com/sites/nizangpackin/2023/05/22/automations-hidden-costs-the-case-against-a-paywalled-human-touch/?sh=4ddef40c3402

Goldberg, S. C. (2012). “Epistemic extendedness, testimony, and the epistemology of instrument-based belief”, *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* 15(2): 181-197. doi:10.1080/13869795.2012.670719

Goldberg, S. C. (2016). “Epistemically Engineered Environments”, *Synthese* 197(7): 2783-2802.

Google AI Blog. (2018). “Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone”. Google AI Blog, May 8, 2018. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Graham, P. J. (1997). ‘What is Testimony?’, The Philosophical Quarterly 47, 227–232.

Grice, H. P. (1971). “Intention and Uncertainty”, *Proceedings of the British Academy* 5: 263-279.

Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90. https://doi.org/10.1016/j.ebiom.2023.104512

Hastings Center Report, 51(3), 17-22. https://doi.org/10.1002/hast.1207

Hatherley, J. J. (2020). Limits of trust in medical AI. Journal of Medical Ethics, 46(7), 478-481. https://doi.org/10.1136/medethics-2019-105935

Hurst, L. (2022). “ChatGPT: Why the human-like AI chatbot suddenly has everyone talking”, EuroNews, December 14, 2022. https://www.euronews.com/next/2022/12/14/chatgpt-why-the-human-like-ai-chatbot-suddenly-got-everyone-talking

Jones, K. (1996). “Trust as an Affective Attitude”, *Ethics* 107(1): 4-25.

Kusch, M. (2002). *Knowledge by Agreement: The Programme of Communitarian Epistemology.* Oxford University Press.

Lackey, J. (1999). “Testimonial knowledge and transmission”, *The Philosophical Quarterly* 49(197): 471-90.

Lackey, J. (2006). “It takes two to tango: beyond reductionism and non-reductionism in the epistemology of testimony”, in: *The Epistemology of Testimony*, 160-189. Oxford University Press.

Lackey, J. (2008). *Learning from Words*. Oxford University Press.

Latour, B. (1988). *Science in action: How to follow scientists and engineers through society.* Harvard University Press.

Latour, B., and Woolgar, S. (1986). *Laboratory life*. Princeton University Press.

Lee, K. M. (2008). “Media equation theory”, in: W. Donsbach (ed.), The International Encyclopedia of Communication, Wiley Publishing.

Licklider, J. C. R., and Taylor, R. W. (1968). “The Computer as a Communication Device”, *Science and Technology* 76(2): 1-3.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... and Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123-1130

Lopez-Lira, A., and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619.

Metzinger, T. (2019). Ethics Washing Made in Europe (Der Tagesspiegel, 2019), https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Miller, B. and Record, I. (2013). "Justified Belief in a Digital Age: On the Epistemic Implications of Secret Internet Technologies", *Episteme* 10(2): 117-134.

Miller, B., and Freiman, O. (2021). “Trust and Distributed Epistemic Labor”, In: J. Simon (ed.), *The Routledge Handbook of Trust and Philosophy,* pp. 341-353. Routledge.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. AI and Ethics, 1-31. https://doi.org/10.1007/s43681-023-00289-2

Mollman, S. (2022). “ChatGPT gained 1 million users in under a week. Here’s why the AI chatbot is primed to disrupt search as we know it”, Yahoo! Finance, December 9, 2022. https://finance.yahoo.com/news/chatgpt-gained-1-million-followers-224523258.html

Moon, Y. and Nass, C. (1996). “How ‘real’ are computer personalities? Psychological responses to personality types in human-computer interaction”, *Communication Research* 23(6): 651-674.

Neges (Kletzl), S. (2018). *Instrumentation. A Study in Social Epistemology*. PhD Dissertation, University of Vienna.

Nickel, P. J. (2013a). “Artificial speech and its authors”, *Minds and Machines* 23(4): 489-502.

Nickel, P. J. (2013b). "Trust in Technological Systems", in: De Vries, M. J., Hansson, S. O., and Meijers, A. W. (eds.). Norms in technology, pp. 223-237. Springer.

Nieva, R. (2018). Exclusive: Google's Duplex could make Assistant the most lifelike AI yet. CNET News, May 9, 2018. https://www.cnet.com/news/google-assistant-duplex-at-io-could-become-the-most-lifelike-ai-voice-assistant-yet

Nissenbaum, H. (1996). “Accountability in a computerized society”, *Science and Engineering Ethics* 2(1): 25-42.

OpenAI. (2022). ChatGPT: Optimizing Language Models for Dialogue. November 30, 2022. https://openai.com/blog/chatgpt/

Origgi, G. (2008). *Qu’est-ce que la confiance*? Paris: VRIN.

Pitt, J. C. (1983). “The Epistemological Engine”, *Philosophica* 32(2): 77-95.

Pitt, J. C. (2010). “It's not about technology”, *Knowledge, Technology and Policy* 23(3-4):445-454.

Pritchard, D. (2004) 'The Epistemology of Testimony', Philosophical Issues, vol. 14, no. 1, pp. 326-348. https://doi.org/10.1111/j.1533-6077.2004.00033.x

Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and Ethical Considerations. In AICS (pp. 104-115). https://ceur-ws.org/Vol-2563/aics_12.pdf

Salinga, M. and Wuttig, M. (2011). “Phase-Change Memories on a Diet”, *Science* 332: 543.

Setiya, K. (2018). “Intention”, in: E. N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/intention

Shanahan, M. (2022). Talking about large language models. arXiv preprint arXiv:2212.03551

Shapin, S. (1994). *A social history of truth.* University of Chicago Press.

Shapin, S., and Schaffer, S. (1985). *Leviathan and the air-pump*. Princeton University Press.

Simon, J. (2015). “Distributed epistemic responsibility in a hyperconnected era”, in: L. Floridi (ed.), *The Onlife Manifesto*, pp. 145-159. Springer.

Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., and Shenefiel, C. (2021). Many hands make many fingers to point: challenges in creating accountable AI. *AI & Society*, 1-13.

Tiku, N. (2022). The Google engineer who thinks the company’s AI has come to life. The Washington Post, 11. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Turing, A. M. (1950). “Computing machinery and intelligence”, *Mind* 59(236): 433-460.

Véliz, C. (2023). Chatbots shouldn’t use emojis. *Nature* 615, 375. doi: https://doi.org/10.1038/d41586-023-00758-y

VoiceBot AI. (2019). *Voice Assistant Consumer Adoption Report 2018*. https://voicebot.ai/voice-assistant-consumer-adoption-report-2018

Waddell, K. (2019). “Defending against audio deepfakes before it's too late”, *Axios*. 3 April 2019. https://www.axios.com/deepfake-audio-ai-impersonators-f736a8fc-162e-47f0-a582-e5eb8b8262ff.html

Wheeler, B. (2020). Reliabilism and the Testimony of Robots. *Techné* 24:2.

Yee, A. K. (2023). Information Deprivation and Democratic Engagement. *Philosophy of Science*, 1-10.