This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2).
(submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

Taking It Not at Face Value:
A New Taxonomy for the Beliefs Acquired from Conversational AIs

Shun Iizuka[1]

Abstract: One of the central questions in the epistemology of conversational AIs is how to classify the beliefs acquired from them. Two promising candidates are *instrument-based* and *testimony-based beliefs*. However, the category of instrument-based beliefs faces an intrinsic problem, and a challenge arises in its application. On the other hand, relying solely on the category of testimony-based beliefs does not encompass the totality of our practice of using conversational AIs. To address these limitations, I propose a novel classification of beliefs that shifts the focus from the properties of the conversational AIs themselves to how recipients perceive AI output, specifically whether they take the output at face value. The proposed categories are *beliefs by regarding as instruments* and *beliefs by regarding as testifiers*. By using these complementary categories, a more comprehensive understanding of beliefs acquired from conversational AIs can be achieved while avoiding the problems associated with the traditional classification.

Key words: epistemology, conversational AI, testimony, instrument-based belief

1. Introduction

The beliefs we have can be categorized according to how they were acquired. For example, beliefs from perception, beliefs from memory, and beliefs from reason, and so on. Such categorization is often useful in identifying issues specific to each category. For each source of belief, various issues have been discussed, such as the veil of perception, preservationism and generativism regarding knowledge from memory, and the problem of induction. On the other hand, the basic question of which source each belief comes from has been less controversial, at least for mundane beliefs that are not acquired through complex processes such as scientific inquiry. In recent years, however, a number of cases have attracted attention in which the category to which the belief in question belongs is itself controversial: beliefs based on conversational AIs.

[1]Shun Iizuka, Department of Philosophy, Graduate School of Humanities and Sociology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan; shun.iizuka.0213@gmail.com

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

Conversational AIs use natural language to interact with users via text or voice. Conversational AIs, including Apple's Siri and Amazon's Alexa, have gradually made their way into our lives over the past decade. In particular, the release of ChatGPT by OpenAI in late 2022 was a surprise due to its fluency and potential usefulness. Other conversational AIs based on large language models (LLMs) have followed, such as Google's Bard and Microsoft's search service Bing AI in partnership with OpenAI. With these conversational AIs based on LLMs in mind, I discuss the beliefs that can be gained from them.

We may come to believe something through our interaction with conversational AIs. Perhaps we may even justifiably believe or know something.[1] How, then, should we categorize the beliefs that we acquire in this way? One possibility is to create an independent new category, as Ori Freiman suggests (Freiman 2023, Sec. 4). However, at least for now, he only gives a rough sketch of the new category of *technology-based beliefs*. When we think about the sources of beliefs, what we are sometimes concerned with is the conditions for their justification or knowledge status. Indeed, the examples given at the beginning as problems specific to each source-based category all ask about justification or knowledge status. However, there has been no substantive discussion of the conditions for technology-based beliefs to be justified or to qualify as knowledge.

Another possibility is to start by considering an existing category as an approximation. There are two candidates: *instrument-based beliefs* and *testimony-based beliefs*. First, if we look at the fact that conversational AIs are implemented in machines, believing the output of conversational AI is similar to believing that the temperature here now is 26℃ when we see a thermometer pointing to 26℃. Therefore, the category of instrument-based beliefs is a promising candidate. The motivation for choosing the other candidate is easier to understand when comparing the following two stories. In the first, I asked a friend, "Who was the 39th President of the United States?" and he replied, "Jimmy Carter, isn't it?" and I believe his statement; in the second story, I type into ChatGPT, "Who is the 39th President of the United States?" and I believe the output content "It's Jimmy Carter." These are obviously very similar. It is therefore also a natural idea to apply the theory of testimony-based beliefs (cf. Wheeler 2020).

The goal of this paper is to gain a better understanding of beliefs from conversational AIs with reference to previous debates about whether beliefs from conversational AIs are testimony-based or instrument-based. This type of research has primarily focused on whether conversational AIs possess properties that would qualify them as testifiers (Freiman and Millar 2019; Wheeler 2020; Freiman 2023). In contrast, I argue that better understanding can be

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

achieved by adopting a new taxonomy that focuses on how the recipient perceives the output of a conversational AI: *beliefs by regarding as testifiers* and *beliefs by regarding as instruments*. This idea of focusing on the recipient's side stems from Christopher Green's argument that machine testimony is genuine testimony (Green 2008, 3.b.ii). According to Green, when a recipient believes something based on machine output, the machine's output can be considered testimony if the recipient's epistemic state, content, the cognitive abilities exercised, and phenomenology are the same as in the case of human testimony. In contrast, I focus not on the acceptance of machine output as if it were human testimony, but rather on the situations in which human testimony or similar conversational AI output is treated as if it were machine output.

The paper is structured as follows. Section 2 critically considers the application of two categories of beliefs, instrument-based beliefs and testimony-based beliefs, to the cases of obtaining beliefs from conversational AIs. Section 3 first refines the theory of instrument-based beliefs to address one of the problems. In doing so, I focus on the fact that the recipient interprets the output of the instrument. Then, we will see that even in cases where the informant is a human or a conversational AI, interpretation sometimes takes place, i.e., the output is taken not at face value. This suggests that the distinction between instrument-based and testimony-based beliefs does not directly correspond to the cases of instrumentally caused beliefs and the cases of testimonially caused beliefs. In section 4, then, I propose a new distinction: beliefs by regarding as testifiers and beliefs by regarding as instruments. Then, cases of taking at face value can be understood as beliefs by regarding as testifiers, and the cases of taking not at face value can be understood as beliefs by regarding them as instruments. Finally, I consider several reasons why cases of taking not at face value are not merely marginal with respect to the beliefs from conversational AIs.

2. Instrument-based Beliefs and Testimony-based Beliefs

2.1 Instrument-based Beliefs

We know the current temperature by checking which degree scale the thermometer is pointing to, or we know the time of day by looking at a clock. These are typical cases of our instrument-based beliefs. Typically, instrument-based beliefs are understood reductively as being based on a combination of perceptions, inferences and so on (Lehrer 1995; Sosa 2006). Here, I take up Ernest Sosa's theory. According to Sosa, for instrument-based beliefs to be justified, the instrument must actually be reliable in the situation (Sosa 2006, 116–17).[2] In addition, Sosa argues that it is necessary to be able to assume, in a non-arbitrary way, that instruments are

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

reliable. Such assumptions will be non-arbitrary if, for example, they are based on inductive generalization or testimony (Sosa 2006, 118). Specifically, one needs to believe in the reliability of the instruments by observing that they repeatedly produce good results or by hearing from others that using the instruments is a good way to find out the time or temperature.

In applying Sosa's theory of instrument-based justification to cases of conversational AIs, however, at least two problems must be overcome. One is a problem of the theory itself, and the other arises when applying the theory to conversational AI cases. The first problem is raised by Ori Freiman and Boaz Miller (Freiman and Miller 2019, 421–22). They argue that there is no such thing as a perfectly accurate instrument, and thus there is no reliable instrument.[3] For example, the standard second is only an idealized one, and thus even the most accurate atomic clocks cannot represent the standard second with perfect accuracy. Similarly, there is no single thermometer that provides a standard, but better standards are gradually defined on the basis of several thermometers that are considered to be the best. Thus, the instruments we use in our daily lives are inaccurate and unreliable. If the above is correct, then following Sosa's idea of instrument reliability as a necessary condition for instrument-based justification would mean that much of what we think we justifiably believe would be unjustified. Specifically, when I see a clock pointing to 1:55:6 and I believe that it is now 1:55:6, my belief is supposedly justified, but in fact it is not, because every clock has a slight error from standard second and is not reliable.

Even if this problem regarding instrument-based beliefs themselves can be overcome, there is a further concern in applying it to the cases of conversational AIs (Freiman 2023, 7). When considering instrument-based beliefs, what we have in mind are primarily mechanical instruments such as clocks and thermometers. The outputs of such instruments are not usually expressed in natural language. Therefore, as in Sosa's theory, instrument-based justification is explained in terms of the reliability of the mechanism of the instrument, our perceptual ability to grasp the outputs, the process of inductive reasoning, and so on. The propositional content of the beliefs we acquire by using the instrument is then left out of the consideration. In contrast, conversational AIs produce natural language outputs based on data and algorithms. Freiman is therefore concerned about applying the category of instrument-based beliefs, which cannot deal with the propositional content expressed by language, to conversational AI cases.

2.2 Testimony-based beliefs

There are different views on the conditions for testimonial knowledge or justification, such as the reductionism/anti-reductionism and transmission/generation debates. However, there

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

seems to be general agreement that for the phenomenon of testimony to hold, the testifier must possess some characteristics that qualify her as a testifier. The relevant characteristics vary from belief (proponents of the transmission view), intention (Hinchman 2005, 567) or responsibility (Goldberg 2012). As long as the standard is based on such characteristics, it is often concluded that machines cannot be testifiers (Lackey 2008, 189; Fricker 2015, 178–79). Conversational AIs, which are still considered to lack the required characteristics, would be no exception. Thus, it is argued that the theory of testimony-based beliefs is not applicable to the cases of conversational AIs (Freiman 2023, 8–10).

One possible response to this concern is to argue that the characteristic required to be a testifier is not such as to exclude at least all conversational AIs. Billy Wheeler (2020, 343–51) argues that the criterion consists in the ability to deceive or withhold information depending on the situation rather than the above characteristics. It is argued that some machines can be testifiers because they meet this criterion. And this makes it possible to apply the reliabilist theory of testimony to the cases of beliefs from machines (Wheeler 2020, 335). According to this theory, justified beliefs based on machine testimony can only be obtained if the machine is reliable. I leave aside here the plausibility of the claim that the relevant property is the ability to behave deceptively, and that some machines have this ability. At the very least, it is doubtful that such an ability is implemented in conversational AIs such as ChatGPT, which is widely available today. However, it seems reasonable to assume that certain machines are quasi-testifiers, and to apply theories of testimonial knowledge, as Wheeler does. This is because the state of mind or responsibility of the testifier is sometimes not seen as a necessary condition for testimonial knowledge.[4] Lackey, for example, requires that the testimony be reliable as a necessary condition for the hearer's testimonial knowledge, while not requiring the testifier's knowledge (Lackey 2006, 47–65, 177–78).

But does a conversational AI necessarily have to be highly reliable for us to use it to justifiably believe something? As is well known, it has been pointed out that conversational AIs based on LLMs often experience *hallucinations*, i.e., the output of sentences that are grammatically correct but are nonsensical or contain falsehoods. Indeed, *misinformation* is counted among the risks of language models (Weidinger *et al*. 2022, 2.3). Given this, it is not clear that current conversational AIs are reliable enough to provide justification or knowledge.

Nevertheless, we are learning to make use of conversational AIs even now. Of course, the immediate purposes are not limited to obtaining knowledge-that such as "the 39th President of the United States is Jimmy Carter." For example, when we ask ChatGPT to suggest a slogan for a new product or to advise us on the plot of a novel, we do not require factivity. However,

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

there are cases where correctness is required even when we are not seeking knowledge-that. Justin Weinberg, for example, suggests some academic uses of ChatGPT and related applications that depend on it, such as helping to understand articles, translating, and paraphrasing one's own writings (Weinberg 2023). Even for such uses, however, in order to use the acquired text, one must be able to reasonably believe, to some degree, that this translation obtained using conversational AI is appropriate, that this paraphrase is intended, and so on. Is the appropriateness of such practices then limited by the reliability of conversational AI?

One promising idea is to assume the ability of recipients to filter information. In his talk, Adam Carter, drawing on Peter Graham (2010), was optimistic and anti-reductionist about the knowledge derived from ChatGPT, saying that the users can filter the propositions that it outputs (Carter 2023). I would like to use a very simple model to show that, when filtering is taken into account, it is possible that we obtain beliefs that are likely to be true even from not very reliable sources of information. Suppose a conversational AI outputs a proposition p and the recipient either believes p or ignores the output. The AI outputs a true proposition with probability a, and the recipient's filtering competence, i.e., the probability that the recipient can correctly discern a true proposition as true, and a false proposition as false, is b. In other words, the recipient believes the AI's output with probability b when p is true, and ignores it with probability $1-b$. He also believes with probability $1-b$ when p is false, and ignores it with probability b. In reality, the filtering ability should not respond to the truth itself, but to other factors that correlated with it, but for simplicity, the above assumptions are made here. Let B be the event that the recipient believes p and T be the event that p is true, then the probability that the recipient's belief that p is true can be expressed as $P(T|B)$. According to Bayes' theorem,

$$P(T|B) = P(B|T)P(T)/P(B).$$

Since the probability that the recipient believes p when p is true, represented by $P(B|T)$, is equal to the filtering competence,

$$P(B|T) = b.$$

Since the probability that the proposition p output by the AI is true, represented by $P(T)$,

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2).
(submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

equals the reliability of the AI,

$$P(T) = a.$$

Since the probability that the recipient believes p, represented by $P(B)$, is the sum of the probability that p is true and the recipient believes it correctly, which equals ab, and the probability that p is false and the recipient wrongly believes it, which is $(1-a)(1-b)$,

$$P(B) = ab + (1-a)(1-b).$$

Substituting these into the above formula obtained by Bayes' theorem, we get,

$$P(T|B) = P(B|T)P(T)/P(B) = ab/(ab + (1-a)(1-b)).$$

Then, for example, if the reliability of the conversational AI, a, is 60% and the filtering competence of the recipient, b, is 90%, then

$$P(T|B) = 0.6 \times 0.9/\big((0.6 \times 0.9) + (0.4 \times 0.1)\big) \approx 0.931.$$

That is, the probability that a belief based on the AI is true is greater than 93%. Thus, if our filtering functions are sufficiently effective, we may be able to use conversational AIs to form beliefs in an appropriate way, even if they are not very reliable.

Of course, I do not think this is a conclusive argument that we can acquire justified beliefs through conversational AIs. Success or failure depends on the actual reliability of AIs and the competence of our filtering abilities. But we are not yet at a point where we can make that judgment. We do not know whether AIs signal falsehoods as well as humans do. If they signal differently, it is unknown whether our filtering abilities are adequate to detect their signals. Moreover, even if we have, or could in the future acquire, filtering abilities suitable for detecting signals that AIs might give, it is questionable whether the abilities will work in AI cases given the ELIZA effect, a phenomenon in which people mistakenly project human characteristics onto a computer program; unconsciously perceiving AIs as if they were humans would prevent us from correctly applying the proper filters for AIs.[5] The actual effectiveness of our filtering abilities in AI cases can only be determined empirically in the future with these

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

considerations in mind.

Can we then understand the entirety of obtained beliefs from conversational AIs as testimony-based? If so, there does not seem to be much we can say at this point about the justification status of such beliefs. However, what we may be doing when we use conversational AIs like ChatGPT is not simply believing or ignoring the output of p. As we will see in the next section, we seem to be using them in more complex ways.

3. Taking It Not at Face Value

In this section, I first refine the category of instrument-based beliefs by examining the intrinsic problem raised by Freiman and Miller. I then present several cases of taking not at face value that do not fit well into the existing taxonomy of instrument-based/testimony-based beliefs.

3.1 Responding to Freiman and Miller's Criticism

Sosa requires instrument reliability as a necessary condition for instrument-based justification, but even the best instruments are not perfectly reliable, making any instrument-based justification impossible. For example, it would be impossible to justifiably believe that it is now 1:55:6 by seeing a clock pointing to 1:55:6. This is the gist of Freiman and Miller's critique. But, more precisely, is what I believe and am supposed to know when I see a clock pointing to 1:55:6 really "it is now 1:55:6?" What does it matter if the actual time is 1:55:6.31? The problem arises from the idea that the clock is asserting that it is 1:55:6, and we believe exactly what it says. It is not the extremely precise time that we usually try to find out by looking at a clock. In fact, what I believe should be "it is approximately 1:55:6" or "it is approximately 1:55," or perhaps "it is approximately 2:00." If so, it would seem that the clock, even with some error in the acceptable range, provides knowledge or sufficient justification for our belief. The negligible error of the instrument that Freiman and Miller point out is harmless to our justified beliefs.

The lesson to be drawn from their critique and my response is that when we believe something on the basis of an instrument whose outputs are not expressed in language, we are always making some interpretation, in other words, processing the output information. Since typical instruments do not represent propositions, it is clear that the recipient must process the output information into a proposition to be believed in order to acquire knowledge-that. In the following example, the processing is even more obvious. I always set my house clock 10 minutes ahead to avoid being late. As I am rushing to get ready, I look at my clock and see that it reads 1:55. I know that the clock has been advanced by 10 minutes. So I also know that it is

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

1:45 (approximately, of course), and I am relieved. I am clearly processing information in this case, and it is also epistemically (though perhaps not pragmatically) appropriate. This example may seem to be only an exception, but there is always some interpretation or processing going on. Even when we look at ordinary clocks, we are processing according to a widely shared method.

In light of the above, although Freiman and Miller's challenge is not critical, Sosa's theory is deficient in a strict sense. Since the output of an ordinary clock or thermometer is not in the form of a proposition and is therefore neither true nor false in itself, we cannot expect such instruments to be reliable or to have a tendency to output the truth. What we should require instead is stability, that is, a constant output related to the way the world is. We can gain knowledge from a clock that is always 10 minutes ahead, but we cannot learn anything from an unstable clock that is sometimes five minutes behind and sometimes four hours ahead. Instrument-based justification holds only if the instrument is stable, the recipient believes this properly and processes the output into a belief in a reliable way.

3.2 Taking it not at face value: cases of human testimony

Is it only in the cases of instruments that we process information to form beliefs? In the typical case of beliefs from human testimony, the testifier testifies that p and the hearer believes p on the basis of that testimony. These are the situations we have in mind in the epistemology of testimony. But can we think of a situation in which we process human testimony?

As a counterexample to the transmission view, Graham (2000, 379–80) describes an alien named Alan, who is indistinguishable from earthlings, who speaks the same language as English except that all the vocabulary related to color is inverted, and who comes to Earth from his home planet where all colors are inverted, unknowingly equipped with lenses in his eyes that invert all colors. Because he wears color-reversing lenses, the sky on Earth looks to him like what we would call "yellow," but his color vocabulary is inverted, and so he says, "the sky is blue," as we do. It seems possible for us to know that the sky is blue on the basis of his testimony. Now consider the following counterpart of Alan. He is indistinguishable from earthlings and speaks English just like earthlings. He has no color-reversing lenses in his eyes. He does, however, have a distinct color perception, and all colors appear reversed from those of standard earthlings. The clear sky on Earth appears yellow to him, so when asked what color the sky is on Earth, he replies, "the sky is yellow." And suppose a researcher has been observing him for some time and is aware of the characteristics of his testimony about colors. She can then hear his testimony that the sky is yellow and believe that it is blue. And this belief of hers

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

seems to be justified. In a way, she herself acts as a substitute for the lenses.

The above is a science fictional example, but more realistic stories are also possible. Let us recall a famous anecdote. Just before the Great Depression, investor Joseph Kennedy heard from a shoeshine boy that he could make money by buying stocks. He then judged that the stock price would fall in the near future and decided to sell his shares to avoid a loss. Aside from the veracity of this anecdote, Kennedy in this story used the boy's testimony to form a belief but did not take it at face value. While we cannot say that he knew about the stock market crash, it is possible that his belief was somewhat justified.

3.3 Taking it not at face value: cases of conversational AI outputs

As described above, it is possible to form beliefs from human testimony without taking it at face value. However, such cases are only marginal cases of testimony-based beliefs. It is plausible that the epistemology of testimony has taken as a typical example the situation where a hearer comes to believe p based on the testimony that p. What then about the cases of beliefs from conversational AIs?

First, I would like to highlight translation, also mentioned by Weinberg, as a promising area of application area for conversational AIs based on LLMs. Translation using LLM-based services, including conversational AIs, often suffers from poor translation of technical terms. For example, the standard Japanese translation of "reliabilism" is "信頼性主義 *shinraisei-shugi*," but when ChatGPT is asked to translate a sentence containing this word, it is sometimes translated as "確証主義 *kakushou-shugi*," "信用主義 *shinyou-shugi*," "信頼主義 *shinrai-shugi*" or "信頼性論 *shinraisei-ron*." As another example, the standard translation of "defeat" is "阻却 *sokyaku*," but more often a translation is chosen that does not capture the nuances in the context of the source language, such as "敗北 *haiboku*" (to lose) or "敗訴 *haiso*" (to lose a lawsuit). Thus, ChatGPT is not a reliable translator of the epistemological terminology in the sense that it does not output a sufficiently high proportion of correct translations. Therefore, its output cannot be trusted as a correct translation as it is. Nevertheless, there is still room for ChatGPT to be used for translation. After many attempts at translation in the same field, patterns of errors may emerge, as in the example above. Then we can replace "確証主義

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2).
(submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

kakusho-shugi" in the output translation with "信頼性主義 *shinraisei-shugi*" (reliabilism) and "敗北 *haiboku*" with "阻却 *sokyaku*" (defeat) by ourselves. This is just like the researcher considering what Alan's counterpart suggests by saying "yellow" is "blue" thanks to long-term observation.

Problems with technical terms also arise in the case of proofreading, another promising application of conversational AI. Technical terms should normally be used consistently throughout a document. In English, at the same time, repetition of other common vocabulary is discouraged, and it is recommended to rephrase using different expressions. This gap causes problems. Technical terms are not used consistently and are replaced by similar expressions. This is often a problem, especially when proofreading philosophical texts, where everyday vocabulary is often borrowed as technical terms. However, it is still possible to guess the correct vocabulary if one is aware of these characteristics and of the patterns of paraphrasing for each term.

More generally, other problems have been observed when using services based on LLMs to transform sentences, such as translation, proofreading, and paraphrasing. For example, they sometimes convert sentences containing complex negations or comparisons to their opposite meanings, ignore an entire sentence, and duplicate the same sentence multiple times. However, being aware of these quirks should allow for more correct sentence construction based on the outputs.

The above cases of belief formation without taking human testimony or the output of a conversational AI at face value do not fit well into the category of testimony-based beliefs, which assumes a situation in which a recipient believes p based on the testimony that p. In particular, even though the beliefs in the cases listed in 3.2 are clearly caused by the testimony, they cannot be analyzed as testimony-based beliefs. In other words, the distinction between instrument-based and testimony-based belief does not directly correspond to the cases of testimonially caused beliefs and instrumentally caused beliefs.[6]

4. Beliefs by Regarding as Instruments and Beliefs by Regarding as Testifiers

Therefore, I would like to propose a new taxonomy of *beliefs by regarding as testifiers* and *beliefs by regarding as instruments*. The knowledge and justification conditions for each category are the same as for instrument-based beliefs and testimony-based beliefs, respectively. However, this distinction does not depend on what the source of information is. Instead, it is

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2).
(submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

determined by whether the recipient takes the information at face value, viewing the source as a testifier seeking to tell the truth, or takes it not at face value, viewing the source simply as an instrument reacting to the way the world is.

In the new taxonomy, the cases presented in sections 3.2 and 3.3 are classified as beliefs by regarding as instruments. For a belief by regarding as instrument to be justified, like an instrument-based belief, the instrument must have high stability, the recipient must properly believe in its stability, and he must be able to form beliefs from the output in a reliable way. For example, in the case of Alan's counterpart, his color vision is systematically reversed, and the researcher is aware of this, so she can gain knowledge about the color of things from his testimony. Kennedy's anecdote would be more subtle. The degree to which his belief in a stock market crash is justified depends on whether a person in the position of a shoeshine boy's interest in stocks is really an indicator of a falling stock market, and how justifiably he believed that it was an indicator.

ChatGPT's translation quirks are not as simple as Alan's counterpart, but they are also not so complex as to be completely incomprehensible in most cases. The more we know about the quirks, the more appropriate translations we can get with the help of ChatGPT. In other words, we will be able to get translations that we can justifiably believe are appropriate. Perhaps, even in the cases other than translation tasks, it will also be true that more knowledge about quirks will make ChatGPT more useful as a source of knowledge and justified beliefs. Since ChatGPT was made public, users have been examining and sharing what it is good at and what it is not, what prompts we can type to get the desired outputs, what patterns of errors are found, and so on. They are not just checking whether ChatGPT is reliable or not. Rather, they are trying to get something useful out of ChatGPT through a process of trial and error, while accepting that errors will often occur.

The new taxonomy alleviates the problem of applying the theory of instrument-based beliefs. The problem with applying the category of instrument-based beliefs has been that it is not inherently a concept for instruments whose outputs are expressed in natural language and have propositional content. However, this problem does not arise when we consider cases of taking not at face value as beliefs by regarding as instruments. This is because the propositional content in question is not acquired by a conversational AI itself, but by the recipient. The output of the informant regarded as an instrument, whether it has propositional content or not, is only one of the materials from which the user forms beliefs.

The new taxonomy also helps us go beyond the limitations of testimony-based beliefs. The cases in which the conversational AI outputs a proposition p and the recipient believes p

This is a preprint manuscript of the paper accepted for publication in *Techné: Research in Philosophy and Technology* to appear in 28(2). (submitted August 1, 2023; revised November 17, 2023; accepted January 8, 2024; uploaded April 11, 2024; copyedited version uploaded May 27, 2024; proofread version uploaded July 30, 2024)

by taking it at face value can be considered as a belief by regarding as a testifier. In such cases, whether we can acquire knowledge or justified beliefs depends on the actual reliability of the AI and the effectiveness of our filtering abilities. In addition to this, by allowing room for beliefs by regarding as instrument, we can accommodate cases of taking not at face value, in which the recipient forms a belief other than p from the output of p.

In proposing my new taxonomy, I have emphasized the cases of taking not at face value. One might object that these are only marginal cases. This would certainly be true if only human testimony were considered. However, for several reasons, the outputs of conversational AI are easier to use than human testimony without taking at face value. First, it can be a dishonest attitude to listen to human testimony without expecting it to be true, and then make use of it to believe other propositions arbitrarily. This is because it treats the testifier as if he were an instrument that regularly reacts to the way the world is but did not speak the truth. Such a stance would mean that the testifier is not regarded as a peer epistemic agent to be respected. This is not a problem in the cases of conversational AIs. This is because the AIs that currently exist do not seem to deserve the same respect as humans as a single epistemic subject. At the very least, there would be much less psychological resistance to dishonesty than there is with humans. Conversational AIs can also be repeatedly prompted with slight variations to test their responses. This allows us to learn their quirks. It would be unlikely to do this with a live person, as it would be too rude. More importantly, a single conversational AI can be used by many people and its quirks can be shared. It would be impossible for a single person to communicate with the equivalent of over 100 million ChatGPT users, let alone have his quirks shared widely. It would also require extensive research to find any kind of common quirks in a group of people with certain characteristics. In contrast, conversational AIs, where the same service can be used by many people around the world, allow people to try different prompts and share the quirks they notice. This means that there are many ways to learn reliable ways to derive propositions to believe from their outputs.

5. Conclusion

The traditional problem in the epistemology of conversational AIs has been "Are beliefs from Conversational AIs Instrument-based or Testimony-based?" However, I have not answered this question directly. This is because the taxonomy of instrument-based and testimony-based beliefs is itself unsatisfactory as applied to real cases in which we use conversational AIs.

Specifically, there are two problems with the category of instrument-based beliefs: first, there is the inherent problem that any such belief cannot be justified in a strict sense, and

second, there is the problem in its application that it is not suitable for the cases of conversational AIs with natural language output. The first problem can be addressed by improving the theory of justification of instrument-based beliefs, but not the second. The testimony-based belief category assumes cases where the output is that p and the recipient believes that p. The justification of beliefs in such cases depends on the reliability of the informant and the filtering ability of the recipient, but it is not clear at present whether these are at a sufficiently high level in the cases of conversational AIs. Nevertheless, it seems that we are currently learning to make good use of conversational AIs even for some purposes that require factivity. In other words, categories of testimony-based beliefs alone do not capture the whole of our practice.

Instead, I proposed a new taxonomy of beliefs by regarding as instruments and beliefs by regarding as testifiers. In conclusion, the beliefs from conversational AIs include both of these new categories. If the output is taken at face value, it is the latter; otherwise, it is the former. This way of thinking helps us overcome the problems with the traditional taxonomy.

The new taxonomy reflects the diverse ways in which we use conversational AIs to acquire beliefs. Indeed, it is controversial whether conversational AIs have the relevant properties that qualify them as testifiers, and it is not clear that they are reliable. Nevertheless, we can use different ways of perceiving their outputs and become familiar with their quirks so that we can derive justified beliefs or knowledge from them.

Acknowledgments

I am grateful for helpful discussions with the participants at the Veritas Epistemology Workshop in Seoul, including Peter Graham, Mary Gregg, Masashi Kasaki, Shohei Matsumoto, Nikolaj Pedersen, and Ryo Tanaka. I would also like to extend my thanks to the anonymous reviewers for their comments, which not only improved the readability of the paper, but also allowed me to correct some misleading expressions in advance.

ChatGPT [https://chat.openai.com/], DeepL Translator [https://www.deepl.com/translator], and DeepL Write [https://www.deepl.com/write] were used to improve the English expressions in this paper.

Statements and Declarations

This work was supported by Japan Society for the Promotion of Science KAKENHI Grant Number JP19J21115 and Fuse Academic Foundation.

The author has no competing interests to declare that are relevant to the content of this

article.

References

Carter, Adam. 2023. "ChatGPT and Testimonial Anti-reductionism." Presented at 11th Online Meeting of the Asian Epistemology Network. April 21, 2023 (unpublished).

Freiman, Ori. 2023. "Analysis of Beliefs Acquired from a Conversational AI: Instruments-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs." *Episteme* (First View): 1–17, DOI: 10.1017/epi.2023.12

Freiman, Ori, and Boaz Miller. 2019. "Can Artificial Entities Assert?" In *The Oxford Handbook of Assertion*, ed. Sanford Goldberg, 414–34. Oxford: Oxford University Press.

Fricker, Elizabeth. 2015. "How to Make Invidious Distinctions amongst Reliable Testifiers." *Episteme* 12(2): 173–202.

Graham, Peter. 2000. "Conveying Information." *Synthese* 123(3): 365–92.

Graham, Peter. 2010. "Testimonial Entitlement and the Function of Comprehension." In *Social Epistemology*, ed. Adrian Haddock, Alan Miller, and Duncan Prichard, 148–74. Oxford: Oxford University Press.

Green, Christopher. 2008. "Epistemology of Testimony." *Internet Encyclopedia of Philosophy*. Accessed June 27, 2023. https://iep.utm.edu/ep-testi/

Goldberg, Sanford. 2012. "Epistemic Extendedness, Testimony, and the Epistemology of Instrument-based Belief." *Philosophical Explorations* 15(2): 181–97.

Hinchman, Edward. 2005. "Telling as Inviting to Trust." *Philosophy and Phenomenological Research* 70(3): 562–87.

Lackey, Jennifer. 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford: Oxford University Press.

Lehrer, Keith. 1995. "Knowledge and the Trustworthiness of Instruments." *The Monist* 78(2): 156–70.

Sosa, Ernest. 2006. "Knowledge: Instrumental and Testimonial." In *The Epistemology of Testimony*, ed. Jennifer Lackey and Ernest Sosa, 116–24. Oxford: Oxford University Press.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. "Taxonomy of Risks posed by Language Models." In

Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (*FAccT '22*): 214–29. New York: Association for Computing Machinery.

Weinberg, Justin. 2023. "How Academics Can Make Use of ChatGPT." *DAILY NOUS* February 8, 2023. Accessed June 22, 2023. https://dailynous.com/2023/02/08/how-academics-can-use-chatgpt/

Wheeler, Billy. 2020. "Reliabilism and the Testimony of Robots." *Techné: Research in Philosophy and Technology* 24(3): 332–56.

Notes

[1]I use the terms knowledge and justification quite roughly throughout this paper. This is because they are not so strictly distinguished in existing research discussing beliefs from machines and conversational AIs, and I believe that their differences will not significantly affect my argument. I will therefore proceed by roughly assuming that knowledge requires a considerable degree of justification.

[2]Sosa sees reliability as safety, but here I simply understand it as the tendency of an instrument to produce truth. Also, although both notions of knowledge and justification are used interchangeably in his argument, I here integrate them into justification.

[3]Freiman and Miller do not explicitly distinguish between accuracy and reliability, which are two different concepts. However, since low accuracy implies low reliability, there is no impediment to the argument. I am grateful to Nikolaj Pedersen for pointing out this distinction.

[4]Nevertheless, not all positions on testimonial knowledge or justification are tenable when applied to the machine cases. For example, the transmission view, which holds that the testifier must also have knowledge in order for the hearer to have testimonial knowledge entails that no knowledge can be obtained from machines at all since machines are thought to have no mental states such as beliefs and thus no knowledge that is constitutive of beliefs. This seems to be an unacceptable consequence. However, we find counterexamples such as the creationist teacher (Lackey 2008, 48) to be convincing, and do not consider the transmission view here.

[5]The possibility that the ELIZA effect might influence the exercise of filtering abilities was brought to my attention by the comments of an anonymous reviewer.

[6]The distinction between "based" and "caused" was made clear by comments from Peter Graham.