Should You Trust Your Voice Assistant? It’s Complicated, but No

Filippos Stamatou1 and Xenofon Karakonstantis2

Abstract—The widespread use of voice-assisted applications using artificial intelligence raises questions about the dynamics of trust and reliance on these systems. While users often rely on these applications for help, instances where users face unforeseen risks and heightened challenges have sparked conversations about the importance of fostering trustworthy artificial intelligence. In this paper, we argue that the prevailing narrative of trust and trustworthiness in relation to artificial intelligence, particularly voice assistants, is misconstrued and fundamentally misplaced. Drawing on insights from philosophy and artificial intelligence literature, we contend that artificial intelligence systems do not meet the criteria for participating in a relationship of trust with human users. Instead, a narrative of reliance is more appropriate. However, we investigate the matter further, to explore why the trust/trustworthiness narrative persists, focusing on the unique social dynamics of interactions with voice assistants. We identify factors such as diverse modalities and complexity, social aspects of voice assistants, and issues of uncertainty, assertiveness, and transparency as contributors to the trust narrative. By disentangling these factors, we shed light on the complexities of human-computer interactions and offer insights into the implications for our relationship with artificial intelligence. We advocate for a nuanced understanding of trust and reliance in artificial intelligence systems and provide suggestions for addressing the challenges posed by the dominance of the trust/trustworthiness narrative.

I. INTRODUCTION

Suppose that you fly into a city for the first time. You pick up a rental car and start driving towards your hotel. As the city is unfamiliar to you, you enter the hotel address in the navigation app on your phone and a number of potential routes appear for you to choose. You choose the fastest route, even though it does not follow the main highway into the city. After all, the app has live information on traffic conditions, accidents on the road, or other incidents. The fastest route is the best route, and you trust the app, or at least you think you do. You start driving while the app’s voice assistant is giving you directions: “Turn left”, “In 300 metres, take the first exit” “Continue straight”. Unbeknownst to you, the fastest route takes you right through the city’s most unsafe neighbourhood, where car hijackings are extremely common and where criminals prey on unsuspecting tourists in rental cars, just like you. Fortunately, nothing happens and you arrive safely at your hotel.

But others are less lucky, and they blame the app providers for their misfortunes. In fact, according to a recent news piece, a couple from the US is suing Google because Google Maps led them into a dangerous neighbourhood, where their rental car was attacked and they were seriously injured and robbed of their belongings [1].

What makes users willing to take legal action against a technology company and its navigation app for leading them into an unsafe neighbourhood? Crime is unpredictable, and navigation apps do not claim to protect against hijackers. Yet, users are led to reasonably believe that they can trust their navigation apps in meaningful ways. In this paper, we argue that while this trust is misconstrued and fundamentally misplaced, the reasons for the proliferation of the trust/trustworthiness narrative in relation to artificial intelligence (AI) boil down to how many of these apps are programmed to interact with users. We focus on voice assistants, which are a paradigmatic case of human-machine social interaction and highlight the causes, dangers, and implications of the problematic ambiguity of trust in AI.

We begin by introducing the concepts of trust and trustworthiness in the philosophical literature, as well as selected insights from the growing literature on trust in AI. We conclude that the professed demand for trustworthy AI is problematic, as machines do not meet the criteria for participating in a relationship of trust with a human user. A more appropriate attitude towards AI would be reliance.

Most work on trust and trustworthiness in AI stops here, calling for refocusing the narrative from trust and trustworthiness to reliance. However, we advance the debate by employing the case of voice assistants to explain why the trust/trustworthiness narrative has become dominant. We arrive at a set of factors that enable the perception of a relationship of trust between user and voice assistant. Finally, we argue that this has negative implications for our relationship with AI, now and in the future.

II. TRUST AND TRUSTWORTHINESS

A. Philosophy of Trust and Trustworthiness

Philosophers agree on very little when it comes to trust. However, many converge on the view that trust is an attitude that allows us to depend on others [2] and that it involves risk, usually in the form of vulnerability of the trustor towards the trustee [3]. In contrast, trustworthiness is commonly thought of as a property that a trustee possesses. In this section, we briefly review the literature on trust and trustworthiness in philosophy, both in general and specifically in relation to AI. These insights provide a first approximation of the nature of trust and the conditions under which someone—or something—may be considered trustworthy.

Richard Holton [4] argues that an attitude of trust is unique because it involves taking the participant stance
towards the trustee. In taking the participant stance, the trustor essentially treats the trustee as a person who has the capacity to act freely and be blameworthy or praiseworthy for their actions. Taking the participant stance is the result of considering someone an appropriate target of what P.F. Strawson calls the reactive attitudes [5]. In contrast, if one is not an appropriate target of the reactive attitudes, we adopt the objective stance. For instance, that is how we treat inanimate objects. Importantly, Holton’s theory of trust entails a readiness to feel betrayed by the trustee. Holton claims that the participant stance and the openness to betrayal are absent when we merely rely on someone or something.

Apart from taking the participant stance and accepting the risk of betrayal, trust seems to involve some kind of commitment. Katherine Hawley argues that in trusting others, we expect them to be committed to do what we trust them to do [6], [7]. Believing that the trustee is committed to act in a certain way is important because it enables the trustor to expect certain outcomes and not others. These commitments “can be implicit or explicit, weighty or trivial, conferred by roles and external circumstances, default or acquired, welcome or unwelcome” [6]:11. According to Hawley, to “be trustworthy, in some specific respect, it is enough to behave in accordance with one’s commitment” [6]:16.

While the debate on the nature of trust and trustworthiness is divergent and no single theory emerges as widely accepted, there are a few prerequisites that most theorists agree form part of any good theory of trust: First, the trustor must accept that trusting involves vulnerability. Exactly what one becomes vulnerable to is debatable, but most likely it is to the possibility of some kind of betrayal. Second, the trustee must be willing and able to do what the trustor trusts them to do. Trying to convince someone to do the impossible, or that they will act against their will is not trust in the right sense. In addition, the willingness to act in a way that enables trust should be the result of certain attitudes on the part of the trustee. Finally, in paradigmatic trust relationships, the trustee relies on the trustee to not only hold, but to readily accept any kind of commitment to act according to the trustor’s expectations. Instead, they perform predetermined tasks according to instructions provided to them by a human. In addition, it seems absurd to claim that a human can be vulnerable towards an AI, thereby accepting the possibility that the machine might betray her. So, even with a deflated definition of trust, AI seems unlikely to satisfy the conditions for justifying the attitude of trust in humans or to possess the property of trustworthiness.

Recently, some are calling for a shift in the narrative from trust/trustworthiness to reliability [24], [25], [19]. We largely agree with these calls for refocusing the debate on the reliability of AI systems. Yet, we think that to do that, one must go beyond just pointing out the terminological issue. What is needed is a much better understanding of the unique technical features of specific AI applications (in this case we focus on voice assistants) and the social dimensions of their interactions with humans. This is our focus in the next section. Then, we discuss the implications of the dominance of the trust/trustworthiness narrative in the final section.

III. SOCIAL DYNAMICS OF VOICE ASSISTANT INTERACTIONS

![Image of human-computer interaction model with a voice assistant](image.png)

Interaction with voice-enabled devices has become ubiq-
uitious in the last decade. Voice assistants, including Amazon Alexa, Microsoft Cortana, Google Assistant, and Apple Siri, facilitate various tasks such as information retrieval, meeting scheduling, and hands-free calling from vehicles or homes. Employing Natural Language User Interfaces (NLUI), these assistants engage users and provide services spanning weather updates, navigation guidance, schedule management, and phone call facilitation. Figure 1 illustrates an example of a voice assistant pipeline, which includes language input through speech or text. This process often incorporates voice detection (wake word) models for activation, [26] automatic speech recognition models for converting speech to text, [27] intent recognition models for interpreting user inputs, [28] and text-to-speech models for auditory output. [29] Additionally, cloud computing is commonly leveraged for efficiency, data storage, and edge device utilization, with encryption and supplementary privacy features typically integrated for enhanced security.

In what follows, we touch upon three characteristics of voice assistants that are responsible for both their remarkable success and popularity as well as the idea that they can be trustworthy and humans should be able to trust them. Although these characteristics have already been discussed in relative detail, we suggest how they contribute to the narrative of trust and trustworthiness. Our aim is to disentangle the trust/reliance debate from the complications of mere terminology. We think the demand for trustworthy AI is not just a terminological mistake, but rather rooted in aspects of the technology and its interaction with humans, such as the ones we suggest below.

A. Diverse Modalities and Complexity

Voice assistants, exemplified by Amazon Alexa and Google Assistant, offer a wide range of capabilities that range from mundane tasks such as setting reminders to more complex functions such as offering medical advice. However, this breadth of functionality introduces a level of uncertainty regarding the reliability of each task. For instance, in challenging acoustic environments characterized by high levels of noise, the speech recognition component of these assistants may falter, leading to misinterpretations of user commands or requests. This challenge relates predominantly to the reliability of the system rather than an issue of trust. It is imperative to perform a granular evaluation of each component individually, including the hardware and software responsible for capturing speech signals. This approach ensures a comprehensive understanding of the system’s performance, avoiding reliance solely on instances of misinterpretations by the voice assistant. Additionally, as emphasized by Kim et al. [30], while voice assistants initially captivate users with their intuitive speech-based interfaces, prolonged engagement often reveals difficulties in articulating commands clearly for users, alongside apprehensions regarding privacy and security.

Recent literature has shed light on critical sociotechnical "trust" challenges confronting audiovisual assistive technologies, particularly evident in the domain of hearing aids. [31] These challenges encompass complexities in technology integration, cost constraints, limitations in battery life, and addressing user non-compliance. Similar challenges are encountered in the realm of voice assistants, where the reliability of each aforementioned component is paramount. However, it is unwarranted to attribute untrustworthiness to these technologies solely based on occasional performance limitations. Instead, such limitations underscore the importance of evaluating the reliability of AI assistants across a spectrum of situational contexts. Just as we recognize the multifaceted nature of hearing aids and refrain from questioning their reliability in challenging scenarios, a nuanced understanding of AI assistant technology prompts us to assess their reliability across diverse scenarios, particularly considering their operational capabilities and constraints, i.e., challenges in speech processing, natural language processing, speech synthesis, etc. Therefore, which of these components should we trust?

It is imperative to recognize that characterizing these challenges solely as matters of trust can obscure the crucial distinction between reliability and trustworthiness. This distinction is fundamental for comprehending the factors that shape technology adoption and acceptance within society. A tool is as good as each of its individual parts. Such tools are often also just one component within a broader ecosystem of technologies that aim to improve user experiences in different industries. For example, AI voice assistants are integrated in a wide range of wildly different tasks, from building design tools [32] to helping with the cognitive effort of high-risk decision making [33].

Voice assistants are a paradigmatic case of a highly complex system that performs a wide range of functions, all the while using a simple user interface consisting of just voice inputs and outputs. From the perspective of the human user, a voice assistant can help with diverse everyday tasks, from route navigation to playing music to answering factual questions. And while the inner workings of the system are obscure to the user -and to a certain degree even to the developer-, their usefulness is indisputable.

This combination of complexity and convenience is conducive to a narrative of trust and trustworthiness around voice assistants. While they do not fulfill the conditions for even a deflated notion of trust, the diversity and parallel nature of their functions create an expectation of trustworthiness in the human user. One is hard pressed to pinpoint exactly what a voice assistant is relying on to do. Rather, trusting a voice assistant seems more appropriate, even if the notion is misconstrued.

B. Social Aspects of Voice Assistants

A second aspect of voice assistants that gives rise to a narrative of trust and trustworthiness is their social dimension. That has already been the focus of much work (see for instance [34] for a proposed research agenda). One can ask questions, request information, complete various tasks, receive instructions or navigation, or even have an open-ended conversation with a voice assistant. Importantly, voice
assistants are purposefully designed to promote and maintain intimacy with the user [35]. Research has shown that we are motivated to use voice assistants at least partly by the social benefits they provide [36]. Whether these benefits are real and sustainable or not is debatable [35], but the fact remains that for most people, voice assistants present the first and main opportunity to interact with AI on a daily basis [37].

However, little attention has been paid to the dangers of voice assistants as social interaction partners (see, for instance, [38]). Most voice assistants are programmed to replicate human-like interactions, a feature that often results in inflated anthropomorphization and the attribution of human qualities to these AI systems. For instance, users may opt to alter the accent or gender of their Google voice assistant’s synthesized voice to create a more comforting or familiar experience, particularly when utilizing voice navigation on maps. While this capability can enhance user engagement and satisfaction, it also raises ethical concerns regarding users’ expectations and perceptions. Abercrombie et al. [39] have recently shown that users tend to use gendered (in fact female) pronouns to refer to the most popular voice assistants. At the same time, technology companies routinely deny that their conversational agents are gendered or human-like, even if there is ample evidence to the contrary [40]. Cowan et al. [41] highlight the drawbacks of relying on "humaness" as a metaphor for interaction with voice assistants. The study found that many users viewed the human-like traits of voice assistants as a failed attempt to make them more relatable.

So, at least sometimes, humans interact with voice assistants employing social rules and treating them as social entities. Of course, that is not exclusive; we often treat them as objects. But in many of our interactions with voice assistants, we take the participant stance [42]. This may partly explain why the trust/trustworthiness narrative is particularly prevalent in the literature on voice assistants. In some ways, they are paradigmatic social interaction machines and it is possible to imagine having a meaningful relationship with them. In fact, some are proposing models of creating trust between humans and voice assistants using human-computer interaction theories and para-social relationship theory [43].

C. Uncertainty, Assertiveness, and Transparency

When a voice assistant interacts with a human, it does so typically in the form of assertions, i.e. statements that do not indicate any degree of uncertainty. While event-driven information fusion has potential to convey degrees of confidence by voice assistants [44], users often lack context-sensitive information during interactions with these systems. Presently, there are few mechanisms in place to communicate the confidence level of a voice assistant’s decision-making process to the user. For example, users may not know the probability of a specific road closure due to flash floods in a particularly vulnerable area.

Furthermore, the issue of evaluating absolute accuracy in voice assistants further complicates the matter. Hong et al. [45] conducted a comparison of four major voice assistants’ performance in responding to questions about cancer screening, revealing significant disparities and areas for improvement in the information provided. Their findings suggest issues regarding the reliability of voice assistants, particularly when sourcing responses from the internet without verifying their accuracy. This reliance on potentially inaccurate information may not only undermine what is considered to be the trustworthiness of voice assistants, but also compromise the integrity of the information they provide.

The reliance on internet-derived responses to evaluate the performance of voice assistants clearly raises pertinent concerns regarding their trustworthiness. However, this conflicts with what might be considered as the knowledge base of the voice assistant. Are we subject to trusting the voice assistant or the information available already on the web? This prompts a critical reassessment of the factors that underpin trust in AI systems, emphasizing the pivotal roles of accuracy and reliability in shaping user perceptions. Although matters like transparency remain integral to responsible AI utilization, ensuring the accuracy of information provided by voice assistants takes precedence. Users must trust the reliability of the information conveyed, necessitating the addressing of both epistemic and aleatoric uncertainties by voice assistants.

As mentioned above, transparency is ever so present in the discussion surrounding AI. Schmidt et al. [46] challenge the prevailing notion in explainable AI research that maximal algorithmic transparency inherently fosters trust. Their study underscores instances where transparency, despite its association with AI prediction accuracy, can paradoxically lead to mistrust. This finding prompts a reevaluation of the relationship between transparency and trust within the context of AI systems. Rather than viewing transparency as an absolute solution, it should be regarded as an aspect of responsible AI use. Responsible AI use entails optimizing system performance while meticulously managing confidence levels in predictions (i.e. levels of epistemic uncertainty).

Moreover, the quest for complete transparency may inadvertently create a sense of vulnerability where none should exist. Users rely on the information provided by voice assistants to make informed decisions, and any ambiguity or withholding of information regarding the uncertainty associated with suggestions undermines the reliability of the assistant. This suggests that transparency alone is insufficient to ensure the reliability of AI systems; effective communication of uncertainty is equally crucial. Users should be empowered to navigate the uncertainties inherent in AI systems, enabling them to make informed decisions based on the available information and the assistant’s suggestions. Thus, the reliability of a voice assistant hinges not only on transparency but also on the clarity and completeness of the information conveyed to the user.

IV. NEGATIVE IMPLICATIONS OF THE TRUST NARRATIVE

Now that we have presented a number of factors that contribute to the trust narrative it is important to consider
the stakes. One could reasonably ask "what’s the problem?" Even if the narrative of trust and trustworthiness towards machines is misconstrued and an alternative demand for reliance would be more true to the nature of our relationship with technology, what is the harm?

Let’s revisit the case of the live navigation voice assistant from the beginning of the paper. When the live navigation voice assistant led the unlucky couple into danger, did it fail at being a good navigation app? Probably not. The app is meant to provide route planning and navigation from point A to B, not guarantee a safe passage through the city. Yet, it is plausible that people implicitly rely on the navigation app for other tasks, because the interaction with it is richer than just navigation. It provides alternative routes, informs you about closing times of shops, possible road closings, natural disasters or emergencies in the area, to name a few. It is only natural, then, that people would develop a social-like relationship with their navigation voice assistant, which goes beyond treating it as an object.

A. Ambiguity of Attitudes towards AI

As the case demonstrates, we develop ambiguous attitudes towards AI. On the one hand, we usually treat them as objects, applying to them rules and expectations we normally apply to objects in the world. On the other hand, we sometimes extend to them social expectations and rules normally reserved for humans. These ambiguous attitudes are-at least partly- the result of the three factors we discussed in the previous section. The diverse modalities and complexity of voice assistants, in combination with the wide range of tasks for which humans use them promote a complex relationship between user and machine which goes beyond mere reliance and towards trust. At the same time, the indisputable social benefits of voice assistants make them ideal social interaction partners, moving users naturally towards the participant stance. Finally, the assertiveness and relative opacity of voice assistants restrict the freedom of decision on the part of the user. Since no degree of nuance or uncertainty is revealed, one must either accept or reject what a voice assistant says. A perception of trustworthiness is necessary for choosing to use the technology in the first place.

B. Bias and User Preferences

Apart from the ambiguity of attitudes towards AI, the trust/trustworthiness narrative around voice assistants may also lead to a significant conflict of principles. Because of their strong social dimension, voice assistants learn and adapt to the user’s individual preferences. Over time, a user is more likely to perceive a voice assistant as trustworthy if it becomes more customised upon their beliefs, desires, and even mannerisms. However, most people hold explicit or implicit negative biases against others. So, in the process of catering to individual preferences, a voice assistant runs the risk of producing biased outputs, thus reinforcing negative stereotypes.

C. Mystification of AI

The final implication of the trust/trustworthiness narrative is the mystification of AI. The phenomenon is not exclusive to voice assistants, of course. However, it is important to mention that the mystification of AI often leads to unnecessary anthropomorphism, especially in the form of psychological sophistication. Perhaps the easiest way to substantiate the claim that AI systems can be trusted and trustworthy is by attributing to them agential features otherwise reserved to humans, such as agency, subjectivity, and moral status. Regardless of what one thinks about the possibility of AI agency, it is premature to assign such features to them without considering the implications. For instance, the mystification of AI could be the cause of the phenomenon of responsibility gaps [47], [48].

V. CONCLUSIONS

A. Reflecting on AI Narratives: Shifting Perspectives for Future Development

While our exploration has illuminated various factors influencing the trust narrative surrounding voice-assisted AI while advocating for a shift in the narrative, it is crucial to acknowledge that we are still navigating the evolving landscape of AI technology. We recognize that surrounding concepts like trustworthy AI there is a tendency to personify AI, or a deeply rooted association to cultural pillars such as Isaac Asimov’s "Three Laws of Robotics" which potentially offer thought-provoking ideals for guiding AI development. However, they remain largely in the realm of science fiction due to the current limitations of AI systems. This realization prompts us to engage in discussions that bridge the gap between theoretical aspirations and practical implementation. These discussions should be driven by the diversity in voice assistant models, the social aspect of voice assistants in human-computer interactions and the uncertainty involved in communication with voice assistants or the transparency involved in their use. As we contemplate the future of voice-assisted AI, it is essential to approach these conversations with a blend of optimism and pragmatism, recognizing both the potential and the challenges inherent in the ongoing advancement of AI technology.

REFERENCES

[1] New York Post - Yaron Steinbuch. LA couple claims Google Maps sent them into dangerous South Africa 'Hell Run' area where they were attacked at gunpoint. [Accessed 3 Feb. 2024].


