information

MDPI

Article

Generative AI and Its Implications for Definitions of Trust

Marty J. Wolf [1,*], Frances Grodzinsky [2] and Keith W. Miller [3]

[1] Department of Mathematics and Computer Science, Bemidji State University, Bemidji, MN 56601, USA
[2] School of Computer Science and Engineering, Sacred Heart University, Fairfield, CT 06825, USA; grodzinskyf@sacredheart.edu
[3] Education Sciences and Professional Programs and Deptartment of Computer Science, University of Missouri—St. Louis, St. Louis, MO 63121, USA; keith.w.miller@umsl.edu
* Correspondence: marty.wolf@bemidjistate.edu

Abstract: In this paper, we undertake a critical analysis of how chatbots built on generative artificial intelligence impact assumptions underlying definitions of trust. We engage a particular definition of trust and the object-oriented model of trust that was built upon it and identify how at least four implicit assumptions may no longer hold. Those assumptions include that people generally provide others with a default level of trust, the ability to identify whether the trusted agent is human or artificial, that risk and trust can be readily quantified or categorized, and that there is no expectation of gain by agents engaged in trust relationships. Based on that analysis, we suggest modifications to the definition and model to accommodate the features of generative AI chatbots. Our changes re-emphasize developers' responsibility for the impacts of their AI artifacts, no matter how sophisticated the artifact may be. The changes also reflect that trust relationships are more fraught when participants in such relationships are not confident in identifying the nature of a potential trust partner.

Keywords: trust; e-trust; chatbots; generative artificial intelligence

check for updates

Citation: Wolf, M.J.; Grodzinsky, F.; Miller, K.W. Generative AI and Its Implications for Definitions of Trust. *Information* **2024**, *15*, 542. https://doi.org/10.3390/info15090542

Academic Editors: Herman Tavani and Jeffrey Buechner

Received: 30 July 2024
Revised: 29 August 2024
Accepted: 31 August 2024
Published: 5 September 2024

1. Introduction

In the paper "Why we should have seen that coming: Comments on Microsoft's Tay 'experiment', and wider implications", we examined the case of Tay, the Microsoft AI chatbot that was launched in March 2016. "After less than 24 h, Microsoft shut down the experiment because the chatbot was generating tweets that were judged to be inappropriate since they included racist, sexist, and anti-Semitic language" [1] (p. 54). The case of Tay illustrated a problem with the very nature of learning software (LS) as it was called then: the unpredictability of any software that changes its program in response to its direct interactions with the public will likely cause harm. The Tay incident called attention to the developer's role and responsibility associated with such software and its behavior. The paper made the case that "when LS interacts directly with people or indirectly via social media, the developer has additional ethical responsibilities beyond those of standard software. There is an additional burden of care" [1] (p. 54).

In the eight years since Tay's inelegant debut, much has changed in the landscape of learning software, most dramatically the development of generative AI (GenAI). Data of all sorts are being scraped from the Internet, and algorithms allow software to be changed in response to user inputs [2]. The traditional use of philosophical theories about trust has been challenged by the way information and (more problematically) misinformation are being both used and generated by AI. Big questions have arisen: How much autonomy should chatbots be allowed after deployment? What kind of control should developers retain once a bot is "out of the barn"? What kinds of accountability for the behavior of these automated artifacts are reasonable? In this paper, with an eye toward potentially answering these questions, we step back from them and interrogate our understanding of trust between humans and artificial agents.

In social media, security, education, and business, deepfakes and misinformation are challenging our confidence and trust in technology and in each other. Judges can no longer trust that briefs prepared by attorneys appearing in their courtrooms are citing actual case law. Security vulnerabilities reduce our trust in systems and undermine our trust in communicating with others. We have new issues, but they echo issues from the past: transparency, predictability, identity, and role responsibility, especially regarding algorithms and their use. These new issues call for a re-examination of how trust is understood so that we can better understand and address these problems and the roots of mistrust that seem to be taking hold.

We consider these concerns as we clarify how GenAI chatbots (GenAICs) impact assumptions implicit in definitions and models of trust. In the next section of the paper, we provide a six-facet definition of trust and recount the object-oriented (OO) model of trust that we use. In Section 3, we present case studies and examples of GenAICs and their impact on trust in the context of our definitions. Section 4 provides a point-by-point analysis of the impact that GenAICs have on the assumptions underlying the six-facet definition of trust. The analysis also identifies ways to improve the OO model. Those improvements are presented in Section 5.

We conclude the paper by calling attention to issues surrounding GenAI and GenAIC and lend our support to a comprehensive framework for addressing those issues as they relate to trust.

2. Definitions of Trust and the OO Model of Trust

Because our work on trust involves GenAICs, we start with a definition of artificial agents. An "artificial agent" (AA) is a nonhuman entity that is autonomous (which in this case means that it operates for long stretches of time without human supervision), interacts with its environment, and adapts itself as a function of its internal state and its interaction with the environment [3]. GenAICs are just one example of an AA (see [4] for a comprehensive study on trust in AAs).

The object-oriented model of trust starts with a definition of trust and then uses contextual features to create a model for analyzing trust relationships. For the purposes of this paper, we adopt the six-facet definition used in [5], which is built upon a definition given by [6].

1. Trust is a relation between A (the trustor) and B (the trustee). A and B can be human or artificial.
2. Trust is a decision by A to delegate to B some aspect of importance to A in achieving a goal. We assume that an artificial agent A can make "decisions" (implemented by, for example, IF/THEN/ELSE statements) and that they involve some computation about the probability that B will behave as expected.
3. Trust involves risk; the less information A has about B, the higher the risk, and the more trust is required. This is true for both artificial and human agents. In AAs, we expect that risk and trust are quantified or at least categorized explicitly; in humans, we do not expect that this proportionality is measured with mathematical precision.
4. A has the expectation of gain by trusting B. In AAs, "expectation of gain" may refer to the expectation of the AA's designer, or it may refer to an explicit expression in the source code that identifies this expected gain, or it may be something learned after the AA is deployed.
5. B may or may not be aware that A trusts B. If B is human, circumstances may have prevented B from knowing that A trusts B. The same is true if B is an AA, but there is also some possibility that an AA trustee B may not even be capable of "knowing" in the traditional human sense.
6. Positive outcomes when A trusts B encourage A to continue trusting B. If A is an AA, this cycle of trust—good outcome—more trust could be explicit in the design and implementation of the AA, or it could be implicit in data relationships, as in a neural net.

Fundamentally, the OO model of trust does not depend on starting with this particular definition of trust. Other definitions of trust can be used with the OO model. Once a definition of trust is fixed, contextual features are identified and added to the model. In [5], the two contextual features we used were the nature of the interaction, face-to-face (f2f) trust or electronic (e-trust), and the nature of the participants, human or AA. In the model, the first feature is represented by two subclasses of TRUST: f2f-trust, which requires physical proximity and e-trust, which is established and played out via electronic communication. Figure 1 depicts this superclass/subclass arrangement. It depicts an understanding of TRUST that contains all the features of both e-trust and f2f-trust. Further, there are aspects of e-trust not present in f2f-trust and vice versa.

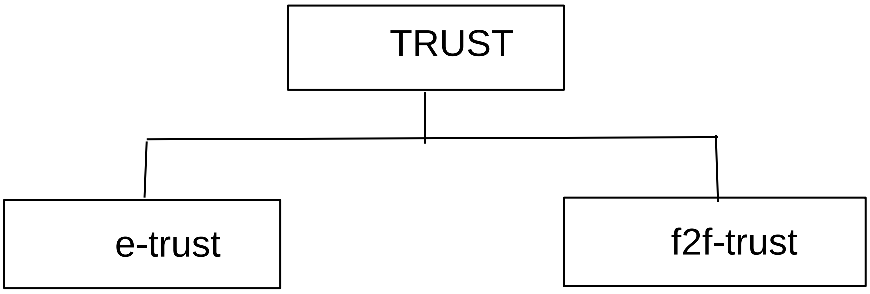

Figure 1. e-trust and f2f trust are subclasses of TRUST.

For the second feature of the model, we identify four classes of interactions: a human trusting a human (H → H), a human trusting an AA (H → AA), an AA trusting a human (AA → H), and an AA trusting an AA (AA → AA). Although a relationship between the same two participants may be initiated and maintained in different modes (for example, I might meet you face to face but later communicate via video conference and email), we handle these complications as mixtures of the "pure" interactions that are found in Figure 1.

Combining Figure 1 and the notion of H and AA trust participants, we have eight new subclasses of TRUST. In Table 1, we only list one-way trust since our underlying definition of trust does not assume that trust is always reciprocal. Note also that f2f is possible with AAs that have a physical presence, such as robots.

Table 1. Eight subclasses of TRUST, four with f2f interactions, and four with electronically-mediated interactions.

H → H f2f-trust	H → H e-trust
H → AA f2f-trust	H → AA e-trust
AA → H f2f-trust	AA → H e-trust
AA → AA f2f-trust	AA → AA e-trust

At this writing, Google Scholar lists 70 citations to [5], which introduced the OO model of trust. We will mention three of the citing articles. In a frequently cited review of the literature exploring the ethics of algorithms, Mittelstadt et al. [7] mention the original model and portray issues of trust in artificial agents as part of a much larger set of issues having to do with trusting algorithms (and software that implements them). We agree with this categorization but choose in this paper to focus on the specific problem of trust involving artificial agents.

Ferrario et al. [8] acknowledge the OO trust model in [5] and build a new model for e-trust by considering three components: simple trust, reflective trust, and paradigmatic trust. In this paper, we also extend the OO model, but in a different fashion than Ferrario et al.

Hou and Jansen [9] point out the importance of shared information and risk when establishing or analyzing trust. Uncertainty about the nature of a potential trust partner is surely a crucial piece of information in this calculation, and our elaboration of the OO model is sensitive to this aspect of trust.

Next, we demonstrate that recent advances in the quality and ease of use of GenAICs are disruptive enough to change our core understanding of trust. So, before proceeding with our analysis, we consider some impacts of GenAI. The examination of these changes drives our re-analysis of our 6 facets of trust, and we identify how the nature of GenAI challenges or contravenes assumptions underlying the facet.

3. Examples of the Impact of GenAICs on Trust

3.1. Education

An important new consideration with the advent of GenAICs is that not only can there be an AA on one or both ends of a communication channel, but a GenAIC can also be an integral part of the communication channel itself. Consider the trust relationship that is built between a particular student and a particular professor. In "ancient times", the student would complete essays by hand. The professor would come to trust that the student was doing their own work because of the physical cues that came from repeated submissions, e.g., similar handwriting, submissions that were generated in class, and in-class interactions. The occasional face-to-face student intent on cheating could still manipulate the trust system, but it was challenging and risky for a student to do so. Contrast this to a course today that is delivered asynchronously via a learning management system, in which students submit type-written essays, and some learning is directed by a GenAIC. Refs. [10–12] all report on using chatbots as teaching assistants.

The H → H relationship that once characterized student/teacher relationships is now often mediated by computers. In some cases, a large part of the student experience has become H → AA, where an AA is doing the grading and some tutoring. In these situations, the trust landscape has changed considerably.

Further, students themselves may well be artificial agents. Courses in general, and online courses in particular, could always be taken by a human student masquerading as another human student. However, AI has decreased the cost and increased the convenience of students getting credit for work they have not done themselves. Student contributions to class, be they in-person discussion contributions, message board postings, or essays, may well be chatbot-generated and merely copied and pasted for the professor to consider [13]. These illustrations demonstrate that all the mechanisms for establishing trust (and distrust) need to be re-examined.

In 2009, a debate arose about a professor who posed as a student in his online class to better understand his students' experiences in his class [14]. Opinions of students and faculty about this practice varied, but many thought this masquerade was a betrayal of student trust. Such a masquerade by a faculty member as a student would be less feasible in a face-to-face class since the effort would be much more elaborate than in an online class. At the very least, this suggests a need to re-consider how the fundamental assumptions about trust have changed.

One important consideration that the student/teacher example brings to the fore is any assumptions that may be present in a model of trust regarding a default level of trust people give others. Here, a professor may be justified in having a default trust level for online students that is lower than that of in-person students. At the very least, it is reasonable for the professor to question whether each student is actually a person. Additional questions arise about whether communications from a student are indeed generated by that student, who intends to convey intent, emotion, and meaning, or whether the communication comes from a vapid chatbot incapable of generating text with an underlying meaning. The professor might be excused for distrusting every student until there is evidence that trust is warranted. Thus, much more time and energy are put into consciously evaluating trust. This stands in stark contrast to "the old days", where trust was the default mode, and one had to notice signs of untrustworthiness in the unusual occurrence that someone was not worthy of that trust. We now have a trust landscape that is less suited for collaboration and more suited for contention.

Questions about how to navigate the altered landscape of trust between students and teachers are outside the scope of this paper. However, educators who are trying to avoid dishonesty while encouraging a climate of trust should consider the significant differences between trust before the rise of GenAIC and trust after and establish classroom policies that address these differences.

The professor/student relationship is merely one example that suggests features that theories of trust need to additionally reflect. We have detailed some ways in which chatbots can impact education and the trust relationships between teachers and students. Though we will not expound on other areas in which chatbots are being utilized, it seems clear that chatbots are now altering, and in the future will alter even more, the trust relationship between buyers and sellers, between people seeking technical help and businesses trying to deliver that help automatically, and between people trying to make appointments and machines programmed to schedule those appointments.

Chatbots are only one set of sociotechnical contexts that utilize new AI technologies. Other contexts share the features of changing trust relationships. Social media is one obvious example. However, legal environments, business environments, and even the process of software development all have changing trust landscapes due to AI. How does the increased vigilance about trust change the social fabric of various contexts? Within groups? Within society as a whole? What are some ways for theories of trust to properly account for these new realities surrounding trust?

3.2. Business and Politics

In a recent incident, an employee at a multinational corporation was tricked into remitting USD twenty-five million. At the time, he was assured by others on the Zoom call that this was a legitimate transaction. Now, it is known that everyone else on the Zoom call was a deepfake recreation [15].

The New Hampshire attorney general's office announced that it was investigating reports of a robocall that apparently used GenAI to mimic U.S. President Joe Biden's voice to discourage voters in the state from going to the polls prior to a recent primary election [16]. Voters who are used to getting calls and seeing ads on TV and social media prior to an election had no reason to believe that this was not a recording of the President speaking to them. These examples illustrate incidents that, in the past, would have been taken as trustworthy. Now, however, with the increasing sophistication of deep fakes, very little can be taken at face value. The more incidents that are revealed as misinformation, the less likely people are going to trust what they see and hear.

3.3. Social Media

Preying on fandom, social media deepfakes have hit communication channels with a vengeance. Some are harmless and created for good, e.g., David Beckham speaking in nine different languages to promote awareness of malaria [17]. Others, however, like the fake images of Taylor Swift supporting Donald Trump at the Grammy Awards and fake nude images of her on X [18], are malicious and damage the credibility and reputation of celebrities, not to mention the psychological harm caused to the person who is the subject of a fake and those receiving fakes.

These examples raise concerns about underlying assumptions in extant models of trust. In the next section, we demonstrate how these concerns manifest in the above definition and offer ways to adapt the OO model.

4. Re-Examining Trust and Its Underlying Assumptions

The considerations from the previous section suggest that the robustness of the assumptions that undergird definitions of trust ought to be re-examined. We undertake this analysis here focusing on our definition of trust adapted from Taddeo and on our subsequent OO model of trust. While our focus is narrow, there is certainly cause for broad consideration of the underlying assumptions for other definitions of trust.

One assumption underlying Taddeo's facet 1 is that the type of agent (human or AA) is knowable by the other agent in the trust relationship. The examples from Section 3 point to situations where that is not always the case. This has implications for the OO model in that one agent in a trust interaction may not know the category of the other. As discussed in the previous section, this epistemic shortcoming complicates the trust-building process and has implications for default trust levels.

A way to codify this epistemic situation is to add an unknown type of agent to the OO model—the uncategorized agent, denoted by a "?". This "agent" is a placeholder for an actual agent whose identity is unknown to the other entity in a potential trusting relationship. Our initial analysis yields that currently in face-to-face situations all agents can categorize the other agent in a trust interaction as either human or artificial (that is, robots today do not pass a physical Turing test); and, thus, there is no justification for adding instances of f2f trust to accommodate this possibility. On the other hand, as the examples in the previous section suggest, there are instances of e-trust that need to accommodate this possibility. This analysis yields four new subclasses of TRUST, as shown in Table 2.

Table 2. Twelve subclasses of TRUST, four with f2f interactions and eight with electronically mediated interactions. We include the first two columns for completeness, but nothing in those columns is affected by the addition of the third column.

H → H f2f-trust	H → H e-trust	H → ? e-trust
H → AA f2f-trust	H → AA e-trust	? → AA e-trust
AA → H f2f-trust	AA → H e-trust	? → H e-trust
AA → AA f2f-trust	AA → AA e-trust	AA → ? e-trust

In our analysis, there is nothing in the nature of GenAI that challenges the underlying assumptions of facet 2, which involves the decision by A to delegate to B some aspect of importance. However, the artificial neurons, especially when taken in aggregate, used in GenAI are more complex than the IF/THEN/ELSE statements mentioned in facet 2.

We call facet 3 into question because of the expectation that in AAs, "risk and trust are quantified or at least categorized explicitly". In GenAI, and by extension GenAICs, trust, or any other attribute of GenAI behavior, typically cannot be quantified or readily identified. GenAI relies on its training and the probabilities embedded in mathematical models. What may appear as "trusting behaviors" may come from mimicking such behaviors that were part of training or may manifest from some learning after deployment. GenAICs, like humans, are less likely to have trust explicitly defined and unlikely to include a number that controls a propensity toward trusting behaviors. To account for this analysis, we offer a revised facet 3.

3. Trust involves risk; the less information A has about B, the higher the risk, and the more trust is required. This is true for both artificial and human agents. We assume that an artificial agent can have behaviors that appear to be "decisions", although it may not have an explicit quantification or explicit programming that can be practically interrogated to examine the nature of those behaviors. We also do not expect that humans can measure their degree of trust with mathematical precision.

Training data plays a significant role in the eventual output of a large language model, which underlies the implementation of GenAICs. The analysis of the role of training data is complicated because of how it is handled. First, raw data are scraped from some source, then curated and cleaned, and then annotated. For our purposes, when we refer to training data, we refer to the raw uncurated and unannotated data. We consider curation, cleaning, and annotation to be part of the development process and under the control of the developer, and we include the people who do that work as "developers". Without explicit tools to identify and characterize information relating to "expectation of gain" in the context of trust, the designers of an AA cannot ensure that their "expectation of gain" manifests in the GenAIC. While it may not be clear that training data add to this "expectation of gain", it is also not clear that they do not. Thus, facet 4 should, at least until there is evidence one way or the other with respect to this issue, be elaborated to include the training data of the AA. Not all AAs are developed using training data. However, it is useful to include consideration of training data in certain trust relationships to emphasize that any trust in an AA that was trained is based in part on an implicit trust in the training data.

The revised facet 4 reads:

4. A has the expectation of gaining by trusting B. In AAs, "expectation of gain" may refer to the expectation of the AA's designer as explicitly expressed in the source code or by information carried in its training data and implemented in its development. It may be something "learned" after the AA is deployed (we use the term "learned" without engaging on the similarities and differences between human learning and GenAI "learning").

Here is a revision of facet 5 emphasizing the increased uncertainty about identities due to GenAI's increasing ability to masquerade as humans:

5. B may or may not be aware that A trusts B. If B is human, circumstances may have prevented B from knowing that A trusts B. Furthermore, a human B may not be certain of whether A is a human or an AA. There is also a difficulty in talking about "awareness" if B is an AA. There is also some possibility that an AA trustee B may not even be capable of "knowing" in the traditional human sense; a GenAI trustee B may not be aware of any individual but is only reacting on the basis of mathematical modeling of behaviors.

Facet 6, we assert, can stand since it already includes mention of neural networks, and GenAICs are based, in part, on their use of artificial neurons.

5. Modifying the OO Model: Do We Need to Create a New Class?

Given the proposed modifications of the definition above, each of the other ten subclasses becomes more complicated. Additionally, the elegance of simplicity in Table 1 is misleading, especially when considering GenAICs. Therefore, we propose a revision of four of the twelve trust subclasses in Table 2. We have not revised the subclasses that only include human-to-human trust relationships for reasons previously explained.

Importantly, there are entities whose impact on trust analysis need to be explicitly considered in a complete analysis of trust involving GenAI. Figure 2 shows these relationships. We note that we have not identified any factors that change trust considerations when a human is the object of trust, so there are no changes to those portions of the model.

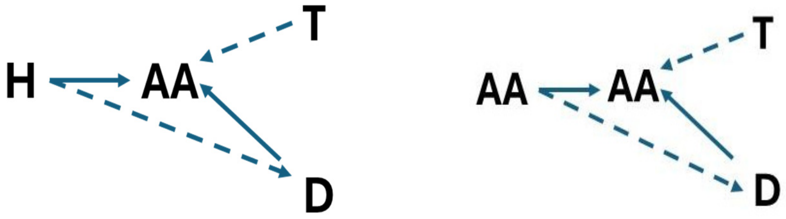

Figure 2. Depictions of the relationships developers (D) and training data (T) have on GenAI trust relationships involving an AA as the object of trust. These relationships hold regardless of whether the trust is f2f-trust or e-trust.

The new symbol D stands for human developers, either as individuals, groups, or corporations, and T stands for training data that were part of the AA's development. We contend that these new features, T and D, are required to appropriately highlight the importance of developers and training when considering whether to trust an AA. We acknowledge this complicates the model, but we are confident that the complications are necessary to avoid hiding important aspects of the trustworthiness of an AA.

The arrow from T is dotted, indicating that not all AAs include training. However, we include it in certain trust relationships to emphasize that any trust in an AA that was trained is based in part on an implicit trust in the training data (see the discussion surrounding facet 4 in Section 3).

The arrow from D to AA is not dotted because there is always human agency that can be traced to the existence of an AA (be it GenAIC or not) and to the design process used. The arrow from a human trustor to the developers is dotted, not because it is optional, but because it is implicit, and the trustor may not be consciously aware of that relationship. Whether the entity trusting is human or AA, the perceived relationship of trust shown in Figure 2 is with an AA. However, we are emphasizing here that the people who develop the AA and the people who embed that AA in a sociotechnical system are involved in that trust relationship, even in cases where they would rather not be involved (these people are represented as D in Figure 2).

In Table 2, we previously introduced the idea of an unknown trust participant. The third column of Table 2 has four relationships involving an unknown participant, marked with a question mark. When the object of a trust relationship is unknown, we do not know which of the depictions in Figure 2 is appropriate. When the source or recipient of a trust relationship is unknown, the resulting depiction just mimics the appropriate depiction for a known participant. So, we do not include any new depictions that include a question mark. We point out, however, that trust when the type of one participant is unknown is riskier than trust with a partner of a known type.

The landscape of GenAI that led to the modifications of our model has necessitated a more finely-grained evaluation of the examples in Section 3. In the education example, the use of GenAI tools has made e-trust more problematic for teachers and students alike. It is more likely now that student work might include the e-trust of an AA. The impact of training sets generating misinformation erodes that trust if the student does not check their work and submits it as is. Submitting such work erodes the trust of the teacher in the student if the work is recognized as compromised.

In the business, politics, and social media examples, humans trust that they are receiving legitimate H–H communications mediated by technology. In reality, this is a case of H-? trust that is misplaced, as the human is trusting the developer, not realizing the output has been manipulated, often by a third party using training data to create a deep fake.

6. Conclusions

The OO model of trust was developed in [5] nearly fifteen years ago. In this paper, we examined the implicit assumptions underlying the definition of trust that were used to develop that model. The examination led to identifying assumptions about trust that do not hold in contexts subject to the use of GenAI and GenAIC. In particular, we found that scholars considering trust should re-examine assumptions about how much trust agents default to, whether there is an assumption that agents can identify whether the trusted agent is human or artificial, whether it is important that, and how, risk and trust can be readily quantified or categorized, and whether an expectation of gain by trusting agents is essential in their analysis of trust. In our case, analyzing the complexity of this trust led us to the expansion of the definition of trust and the classes of the model. Since we are exploring unstated assumptions, it is quite possible that important considerations were not examined. Thus, deeper consideration of how GenAIC, GenAI, and other probabilistic forms of computation contravene unstated assumptions present in analyses of trust and other important philosophical concepts is an essential project going forward.

Further, the sociotechnical system of phones, networks, and corporations collects and exploits enormous amounts of data from individuals that are used in unanticipated ways in many AI applications, including the training of GenAI algorithms (see [19] for a discussion on a phenomenological–social approach to trust). Datasets used in training are a significant concern as they often perpetuate biases and misinformation present in them. The misinformation problem is compounded by the probabilistic nature of GenAI. These observations also call for a re-examination of the implicit assumptions present in philosophical analyses so that they can be refined considering the complexities introduced by these technologies.

For example, it is reasonable to assume that some will deploy GenAICs masquerading as humans increasingly frequently and often with success. In the age of GenAI deepfakes, we have evidence of f2f H → AA interactions where the human is unaware, or at least unsure, if the AA is human or not (see [15,20,21]). Such possibilities may contravene implicit assumptions for many philosophical analyses, including those of trust. Further complicating the analysis is a recent theoretical claim that a GenAIC will be limited to at most fifteen minutes of convincing human behavior [22] (p. 9).

Other factors challenging implicit assumptions are worthy of future study. The emergence of ChatGPT and the subsequent rush to compete with it [23] have resulted in AI artifacts being implemented in many different systems. Some experts in security are alarmed that these systems are coming online far too quickly to assess, let alone mitigate, the vulnerabilities that the AI artifacts are opening [24]. While this is another example of how computing professionals and corporations seem not to be taking seriously their responsibilities to foresee and forestall harms that are possible (and that some would say are likely) when AI artifacts are deployed without sufficient care, it also serves as a challenge for scholars engaging in theoretical analysis of concepts like trust.

In 2020, Randolph Morris, a software development consultant, stated, “The emphasis is almost entirely on getting a product out to market” [25]. That trend has only intensified as AI grabbed the attention of the public and industry. However, this intense rush to market to grab market share means that aspects such as security rarely receive the attention they deserve. Having a model of trust embedded in the development process might move GenAI developers to consider the impact of their products on society and prevent putting out fires after the fact.

Our work here points to important changes to fundamental theoretical models of trust brought about by the nature of and perceived improvement in AI’s effectiveness. Other new or rapidly developing technologies may have similar impacts. Our analysis points to important changes that need to be incorporated into the ethical landscape surrounding trust. Although the public and some computing professionals have been paying far more attention to AI ethics in the last few years [26], the issues laid out years ago (for example, in [3]) are still fundamentally unresolved. Our shifting understanding of trust stemming from GenAI impacts the nature and understanding of the ethical issues surrounding trust in many different segments. Confronting trust issues effectively now reduces the likelihood of harm due to theoretical models that do not properly account for the nature of AI-based systems.

Author Contributions: All authors have contributed equally to conceptualization, formal analysis, investigation, and all aspects of writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wolf, M.J.; Miller, K.; Grodzinsky, F. Why we should have seen that coming: Comments on Microsoft’s Tay experiment, and wider implications. *ACM SIGCAS Comput. Soc.* **2017**, *47*, 54–64. [CrossRef]
2. Orseau, L.; Ring, M. Self-modification and mortality in artificial agents. In *Artificial General Intelligence. AGI 2011*; Lecture Notes in Computer Science; Schmidhuber, J., Thórisson, K.R., Looks, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6830, pp. 1–10.
3. Grodzinsky, F.; Miller, K.; Wolf, M.J. The ethics of designing artificial agents. *Ethics Inf. Technol.* **2008**, *10*, 115–121. [CrossRef]
4. Grodzinsky, F.; Miller, K.; Wolf, M.J. Trust in artificial agents. In *The Routledge Handbook on Trust and Philosophy*; Simon, J., Ed.; Routledge: New York, NY, USA, 2020; pp. 298–312.

5. Grodzinsky, F.; Miller, K.; Wolf, M.J. Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?". *Ethics Inf. Technol.* **2011**, *13*, 17–27. [CrossRef]
6. Taddeo, M. Defining trust and e-trust: From old theories to new problems. *Int. J. Technol. Hum. Interact.* **2009**, *5*, 23–35. [CrossRef]
7. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **2016**, *33*, 2053951716679679. [CrossRef]
8. Ferrario, A.; Loi, M.; Viganò, E. In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philos. Technol.* **2020**, *33*, 523–539. [CrossRef]
9. Hou, F.; Jansen, S. A systematic literature review on trust in the software ecosystem. *Empir. Softw. Eng.* **2023**, *28*, 8. [CrossRef]
10. Chen, Y.; Jensen, S.; Albert, L.J.; Gupta, S.; Lee, T. Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Inf. Syst. Front.* **2023**, *25*, 161–182. [CrossRef]
11. Essel, H.B.; Vlachopoulos, D.; Tachie-Menson, A.; Johnson, E.E.; Baah, P.K. The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *Int. J. Educ. Technol. High. Educ.* **2022**, *19*, 57. [CrossRef]
12. Labadze, L.; Grigolia, M.; Machaidze, L. Role of AI chatbots in education: Systematic literature review. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 56. [CrossRef]
13. Shalby, C. Fake Students Enrolled in Community Colleges. One Bot-Sleuthing Professor Fights Back. LA Times. Available online: https://www.latimes.com/california/story/2021-12-17/fake-student-bots-enrolled-in-community-colleges-one-professor-has-become-a-bot-sleuthing-continues-to-fight-them (accessed on 15 March 2024).
14. Parry, M. Online professors pose as students to encourage real learning. *Chron. High. Educ.* **2009**, *55*, A10.
15. Chen, H.; Magramo, K. Finance Worker Pays Out $25 Million after Video Call with Deepfake 'Chief Financial Officer'. CNN. Available online: https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html (accessed on 15 March 2024).
16. Bohannon, M. Biden. Deepfake Robocall Urging Voters to Skip New Hampshire Primary Traced to Texas Company. Forbes. Available online: https://www.forbes.com/sites/mollybohannon/2024/02/06/biden-deepfake-robocall-urging-voters-to-skip-new-hampshire-primary-traced-to-texas-company/?sh=6c4b5f4b241b (accessed on 15 March 2024).
17. Sodji, L. How We Made David Beckam Speak 9 Languages. Synthesia. Available online: https://www.synthesia.io/post/david-beckham (accessed on 15 June 2024).
18. Tenbarge, K. Taylor Swift Deepfakes on X Falsely Depict Her Supporting Trump. NBC News. Available online: https://www.nbcnews.com/tech/internet/taylor-swift-deepfake-x-falsely-depict-supporting-trump-grammys-flag-rcna137620 (accessed on 15 March 2024).
19. Coeckelbergh, M. Can We Trust Robots? *Ethics Inf. Technol.* **2012**, *14*, 53–60. [CrossRef]
20. Bond, S. AI-Generated Deepfakes Are Moving Fast. Policymakers Can't Keep Up. NPR. Available online: https://www.npr.org/2023/04/27/1172387911/how-can-people-spot-fake-images-created-by-artificial-intelligence (accessed on 5 April 2024).
21. Cai, Z.G.; Haslett, D.A.; Duan, X.; Wang, S.; Pickering, M.J. Does ChatGPT Resemble Humans in Language Use? Available online: https://arxiv.org/abs/2303.08014 (accessed on 15 March 2024).
22. Van Rooij, I.; Guest, O.; Adolfi, F.G.; de Haan, R.; Kolokolova, A.; Rich, P. Reclaiming AI as a theoretical tool for cognitive science. *PsyArXiv* **2023**. [CrossRef]
23. Weise, K.; Metz, C.; Grant, N.; Isaac, M. Inside the A.I. Arms Race that Changed Silicon Valley Forever. The New York Times. Available online: https://www.nytimes.com/2023/12/05/technology/ai-chatgpt-google-meta.html (accessed on 15 March 2024).
24. Wu, X.; Duan, R.; Ni, J. Unveiling security, privacy, and ethical concerns of ChatGPT. *J. Inf. Intell.* **2024**, *2*, 102–115. [CrossRef]
25. Lawson, G. 5 Examples of Ethical Issues in Software Development. TechTarget. Available online: https://www.techtarget.com/searchsoftwarequality/tip/5-examples-of-ethical-issues-in-software-development (accessed on 15 March 2024).
26. Floridi, L. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*; Oxford Academic: Oxford, UK, 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.