Reliability Without Trustworthiness: A Philosophical Exploration of Artificial Intelligence, with a Focus on ChatGPT

This discourse seeks to interrogate the philosophical dynamics of reliability and trustworthiness as they pertain to artificial intelligence, specifically focusing on ChatGPT. The nexus between belief and knowledge is key to the comprehension of reliability, which is essentially predicated on a traceable process that forms a particular belief. Trust, on the other hand, extends beyond this definition to accommodate scenarios where our personal expertise falls short. It embodies a psychological disposition to deem a belief as reliable and assumes the good intentions of the belief's originator. It also implicitly relies on the existence of experts who could verify the belief formation process, if necessary. Therefore, trust is not only the acceptance of reliability but also the faith in the availability of external validation and ethical conduct of the originator.

The metaphor of a medical diagnosis aptly exemplifies this difference. We are often unable to scrutinize the analytical process leading to the diagnosis due to a lack of expertise. However, we accept it based on trust in the physician's expertise, faith in their good intentions, and the comforting knowledge that other medical experts could assess this process if required. When personal capacity to judge reliability is insufficient, we depend on the expertise of others, establishing trustworthiness as a psychological surrogate for reliability.

In a similar vein, we can appraise the reliability of simplistic artifacts like a toaster, given its uncomplicated functionality. However, complex systems, such as computers or AI technologies, necessitate trust due to our limited understanding of their intricate mechanisms.

ChatGPT, in its current state, presents a unique challenge. It is neither reliable nor trustworthy. Both average users and experts find it challenging to evaluate its reliability due to the probabilistic nature of its model. Unlike a conventional deterministic computer program where the designer knows and controls every step, with ChatGPT and similar AI models, the design process does not dictate the exact output. The intention of the designer is also less significant since the probabilistic nature of the model leads to varied outcomes. Thus, the belief formation process remains opaque to both users and its own designers, leaving no experts who can fully understand and validate the process.

I have proposed in my other works that enhancements, such as incorporating perceptual modalities during training and inference phases and assigning credibility-based weights to datasets, could augment ChatGPT's reliability. However, trustworthiness will continue to elude AI systems like ChatGPT. Trust hinges on discerning the good intentions of the system – a fundamentally impossible task in the realm of AI. Probabilistic AI models neither reflect the intent of the designer nor possess the requisite agency and consciousness to foster trust.

In conclusion, users of AI must accept a measure of risk in relying on information produced by such systems and adapt to a paradigm of reliability devoid of accompanying trustworthiness. Future research should explore avenues to alleviate this risk while amplifying the reliability of AI systems.