**So, Can/Should We (Only?) Allocate—Epistemic— Trust to [AI; ML; DNN; LLM; GPT+]?**

In two recent papers, Alvarado argued for two interconnected claims about the nature of AI and our relationship to its role in our knowledge-creating practices. In particular, he argued the following:

1. That artificial intelligence— understood as the domain of technical artifacts that supervene on computationally-intensive and data-science-related processes— is first and foremost an epistemic technology (2023)[[1]](#footnote-1). And,
2. That as such[[2]](#footnote-2), the only kind of trust that should be allocated to AI, if any, is epistemic trust (2022)—the kind of trust allocated in virtue of the recipient’s capacity to store, maintain, manipulate, process, transfer, etc., knowledge-apt content (Simon, 2010; Wilholt, 2013; Alvarado, 2022). [[3]](#footnote-3)

Alvarado, however, notes that “narrowing down the kind of trust that is adequate for the kind of instrument that AI is does not tell us whether the instrument itself is *trustworthy* and hence whether or not our trust [allocation] is well-grounded” (2023 p.13). Thus, he states that “even after it is established that epistemic trust is the only legitimate kind of trust to allocate to epistemic technologies, whether or not AI can in fact be trusted remains an open question” (Ibid p.1). In this paper expand on Alvarado’s framework and explore this question with the aim of articulating some indispensably desirable criteria for such epistemic trust to be appropriately allocated, and the aim of exploring some of the contexts in which AI meets such criteria, where it does not, where it can and how, and where it cannot and why. As we will see, in epistemic contexts such as scientific inquiry, the interrelatedness between trustworthiness and reliability, as well as between transparency and epistemic authority becomes non-trivial (Simon, 2010) such that the epistemic challenges of opacity, the non-possibility of factive explanation in complex computational methods, and the strange/catastrophic error in ML make it so that we may have to contend with the fact that opaque-yet-successful novel computational methods such as AI are not yet trustworthy, epistemically or otherwise.

1. These are, according to Alvarado (2022) technical artifacts that partake in at least two of the following: being deployed in epistemic contexts such as inquiry, being deployed in such a context to manipulate/interact with epistemic content (e.g., propositions, data, databases, books, symbols, sounds), and/or manipulating/interacting with such content through epistemic operations (computing, parsing, labeling, predicting, analyzing, discovering, extracting, etc.)

   [↑](#footnote-ref-1)
2. This is so particularly given that distinct kinds of artifacts should be allocated distinct kinds of trust depending on/respective to their functional capacities (i.e., they should be trusted only for what they can/are/are believed to be able or were designed to do), [↑](#footnote-ref-2)
3. Note too the contrast case in which trusting something for things it cannot/is not designed to do, while possible, seems at best suboptimal, at worst conceptually confused. [↑](#footnote-ref-3)