Trusting Trustworthy AI: what's the right framework?

Feb 2025

Trust – and with it, related notions like trustworthiness – has a long history in philosophy, business, and other settings. In systems thinking, it has at various times been in favour as a counterpart of security. Meanwhile, we see increasingly many systems based on 'AI': and calls for that AI to be trustworthy. Applying ill-defined criteria to an ill-defined class of systems does not appear to be a constructive pastime. But is it constructive to consider a framework for discussing such things?

We propose a hierarchy of kinds of trustworthiness. A foundational layer concerns accurate metadata: in the simplest case, we might see that as a type system, encompassing data sets and schemas as well as the individual datum. Assuring the identity of particular instances may be achieved through signing and hashing. This in turn enables us to make statements about the veracity of AI models, and the data on which they were trained.

A second layer concerns itself with the technologies that have come to be known as trusted execution environments. These allow strong statements about the presence of particular code or data accessible to a particular process: but in order to be useful they must rely upon the foundation layer (so that code and data identities are clear).

Higher layers concern themselves with nuanced semantic properties which some have associated with trustworthy AI: ethics, responsibility, explainability, and so on. Whilst deciding whether a particular system exhibits such properties is not a straightforward task (and depend strongly on careful definitions), it is clear that in order to evaluate such trust meaningfully, the lower levels of the hierarchy are a necessary condition.

This model seeks to integrate prior and current thinking on the topic. The foundation layer incorporates TAIBOM[1], a current project defining schemas to codify a Trustworthy AI Bill of Materials; the thinking is similar in approach to the US AI-SBOM. This targets several use cases for assurance of AI-based systems, including licence validation, security assessment, data flow validation, regulatory compliance, and other risk management.

In this talk, we will describe this hierarchy of models and seek to validate or evaluate their usefulness, in order to attempt to answer the hardy perennial question, which seems to have gained a new currency in the age of 'AI': how shall I decide whether to trust the system in front of me?

---

[1] https://taibom.org