How do we assess system trustworthiness? Introducing the Trustworthiness Assessment Model

2023-07-03 08:00:16

Authors

Nadine Schlicker, Kevin Baum, Alarith Uhde, Sarah Sterz, Martin C. Hirsch, Markus Langer

Abstract

Designing trustworthy AI-based systems and enabling stakeholders to accurately assess the trustworthiness of these systems are crucial objectives. For instance, system developers need to assesses the trustworthiness of systems they are developing before deploying it, auditors need to assess the trustworthiness of a system to grant it market entry, and end users need to assess the trustworthiness of a system for their task goals. Only if trustors (the stakeholder who assesses the trustworthiness of a system) assess system trustworthiness accurately, they can base their trust on adequate expectations about the system and reasonably rely on or reject its outputs. However, the process by which trustors assess a system?s actual trustworthiness to arrive at their perceived trustworthiness remains underexplored. In this presentation, we conceptually distinguish between trust propensity, trustworthiness, trust, and trusting behavior. Drawing on psychological models of assessing other people?s characteristics, we present the two-level Trustworthiness Assessment Model (TrAM). At the micro level, we propose that trustors assess system trustworthiness based on cues associated with the system. Such cues could for instance be single outputs of the system, information about training data, the design of the user interface, or ?trustworthiness labels?. The accuracy of the trustworthiness assessment depends on cue relevance and availability on the system?s side (i.e., which cues are available and are they relevant to assess a system?s actual trustworthiness), and on cue detection and utilization on the human?s side (i.e., what cues does the human detect and how strongly do they utilize those cues to form their perceived trustworthiness). At the macro level, we propose that individual micro-level trustworthiness assessments propagate across different trustors ? one stakeholder?s trustworthiness assessment of a system affects other stakeholder?s trustworthiness assessments of the same system. For instance, auditors may assess a system?s trustworthiness to be high enough to grant the system a

?trustworthiness label?. This is a cue that can then be detected and utilized by other stakeholders in their trustworthiness assessment which may substantially affect their perceived trustworthiness of the system. The TrAM advances existing models of trust (especially Lee & See, 2004 and Mayer et al., 1995) and sheds light on factors influencing the (accuracy of) trustworthiness assessments. It aims to advance conceptual clarity in the area of trust in AI-based systems and has implications for system design (e.g., design systems in a way that makes relevant cues available), stakeholder training (e.g., train stakeholders to detect and adequately utilize relevant cues), and regulation related to trustworthiness assessments (e.g., specify which cues need to be made available).

Keywords

Trust in automation, trustworthiness, trustworthy AI, human-centered design, calibrated trust