THE NECESSITY AND POSSIBILITY OF TRUSTWORTHY AI

Hagen Braun & Lukas Albrecht

Aims of the paper

De-buzzwordify Trustworthy AI

Trustworthy AI seems like a “magic concept” if ever there was one.

Give a theory-driven account of the concept

The EU Assessment List for Trustworthy AI is eclectic. We aim to take the existing scholarship on trust seriously and use it to unify the concept of trustworthy AI.

Explain how trust in the AI context helps bring about better ethics

If the ALTAI list is an ethics guideline, why is this aim expressed in terms of trust?

Necessity of Trustworthy AI

Why do we need trustworthy AI?

Relying on AI is risky **in the same characteristic way** that relying on other (human) agents is risky.

This is because AI is designed to perform **action and decision making** tasks, unlike other technology.

This risk cannot be removed entriely (e.g. through strict control) without losing what makes AI use desirable in the first place.

Frequently, we value the capacity of AI systems to act **autonomously** or in **unfamiliar** environments.

Necessity of Trustworthy AI

The paradigmatic way to manage low, but non-zero amounts of risk in the context of cooperation between agents is through trust. This is the constitutive function of trust.

And AI use can be thought of as **cooperation** between AI and human agents (decision support systems, manned-unmanned teaming…).

Cooperation is generally valuable. It allows cooperants to achieve goals they cannot achieve alone or through just rational self interested modes of action.

But this necessitates a trustworthy cooperative partner.

Therefor, trustworthy AI is necessary.

What is Trust?

Trust (relation)

Trust is a three-place relation between two agents and a task:

T(A_1,A_2,Φ)

The first agent A_1 (the trustor) trusts the second agent A_2 (the trustee) with the task Φ.

With regard to its agential components, this relation is not symmetric, reflexive or transitive.

Important takeaways

This definition of trust implies that we need to trust the *AI agent,* not merely its developers, because the AI agent performs the task.

What is Trust?

That definition alone does not tell us enough. After all, we previously also committed to this:

> **Trust (constitutive function)**
>
> The constitutive function of trust is to manage *agential risk*. Agential risk is the kind of risk that attaches to interactions or cooperation with other agents in virtue of their agential autonomy.

Again, the risk that necessitates trust attaches to the AI agent, not (merely) the developers.

But agential risk is not effectively managed by just *any* kind of trust. If you trust the wrong people they will likely betray you.

So who *should* you trust?

What is Trustworthiness?

Trustworthiness

Trustworthiness is a complex property of agents. It reduces to the properties of *reliability*, *competence* with regard to a certain task, and the possession of a *properly normatively grounded motivation* to perform that task.

Trust is semi-voluntary: An agent can decide to trust anyone they want.

But *well-placed trust* requires identifying what agents *deserve* your trust (i.e.: who is trustworthy).

Only well-placed trust actually manages risk in cooperative relationships.

Preconditions of Trustworthiness

Breaking Down Trustworthiness

Reliability	Competence	Normative Motivation
		

- Reliability is a general precondition for well-placed trust
- Competence makes a person trustworthy with regard to a specific task
- Normative motivation provides an explanatory basis of how trust can manage risk

Are AI Agents Real Agents?

Basic Theory of Action

An action requires a belief-desire pair that constitutes the reason for the action.

It both *causally explains* the action and *justifies* it from the perspective of the agent.

- For AI agents, we can defend the claim that they possess functionally equivalent states to beliefs and desires that can play their causal role.
- But to the best of our knowledge, **AI agents do not possess a perspective from which they justify their own actions in normative terms.**

Are AI Agents Real Agents?

If we take action theory seriously, we cannot categorize AI agents as full agents.

However, from the outside, their behaviour is indistinguishable from real action in many cases.

We therefor want to categorize them as **Pseudo-Agents**: Their actions have the causal component of a real action, without having any internal justifying component. **This means they are still a source of agential risk.**

An AI agent

The Possibility of Trustworthy AI

However full normative agency is not a precondition for trustworthiness.

The possession of a certain kind of motivation for one‘s actions that can be adequately explained in normative terms is.

AI agents can possess motivational equivalent states that are explainable in this way, even though they cannot provide these explanations themself.

But AI developers can.

If we view AI actions as an exercise of joint agency, then the concept of trustworthy AI does make sense. Notably, it does not *just* reduce to trust in AI developers. The agent who acts, and whose agential risk is managed through the trust relationship is still the AI agent.

Trustworthy AI: Final Takeaways

Trustworthy AI is necessary.

AI agents are full-fledged sources of agential risk. Well placed trust is the paradigmatic way to manage this risk.

Trustworthy AI is possible.

AI can be reliable, competent and its motivational equivalent states can be normatively justified.

However, trustworthy AI is hard.

„Normative justification“ in the case of AI motivation requires considerable commitments in terms of ethical standards on the part of AI creators.

And since a part of well-placed trust is correctly tracking when someone is trustworthy, also considerable commitments in terms of transparency.