



Universität Stuttgart

Credible LLMs

The Social Epistemology of AI

Jörg Löschke

The Program

1. Credible LLMs – why?
2. Credible LLMs – how?
3. Social epistemological implications

Social Epistemology

- Epistemology: the study of (the nature, sources, and pursuit of) knowledge.
 - Classic approach: heavily individualistic.
- *Social* epistemology: is concerned with how people pursue the truth, or form beliefs, with the help of other people.
 - What is the best strategy to mitigate the consequences of climate change? Are vaccines safe? Is home schooling harming my child?

Credible LLMs – why?

Large Language Models are increasingly used as *experts*, and not only as sources of knowledge, but also as *advisors* (in medicine, financial matters, etc...)

- Hence, they are increasingly part of our social epistemological practices.
- As such, they must meet the standards of those practices.

Credible LLMs – why?

- Is reliability enough?
 - No: reliability matters as part of a causal chain of events
 - „[W]hen one relies on something to happen one works with the supposition that it will happen into one's plans” (Holton 2007)
 - Advice: not only a causal factor, but a reason-giving factor.
- Social epistemological practices: Two ways of assessing advisors – trustworthy and/or credible.

Trustworthy AI vs. Credible AI

Hypothesis: We should not only strive to make AI trustworthy, but also to make AI systems credible.

Trustworthy AI

- Trustworthiness: understood as „reliability plus X“ (Hawley 2014) (usually understood as goodwill, see Baier 1986).
- Goldberg (2020) on reliance: “where X is a person, artifact, or natural process, and ϕ is an action, behavior or process, to rely on X to ϕ is to act on the supposition that X will ϕ ”.
- Trusting X to give good advice implies the belief that X will give good advice, and that X will do so for specific reasons (such as: caring about the trusting agent for their own sake).
- Epistemic as well as aretaic dimension.

Trustworthy AI

- Problem: AI cannot be trustworthy in this sense, since AI systems do not care about anything and therefore lack goodwill (DeCamp and Tilburt 2019; Fossa 2019; Hatherley 2020; Ryan 2020).
- Does it make sense to talk about trustworthy AI in the first place?
- Yes: In the case of AI systems, reliability is not enough; we also want AI to align with ethical values (Zanotti et al. 2024).
- Alignment with ethical values accounts for the aretaic dimension of trustworthiness.

Why is trustworthy AI not enough?

1. Too encompassing: not all ethical values that make AI trustworthy are relevant for every specific AI system.
 - trustworthiness seems to be a requirement for AI systems in general, rather than individual expert LLMs.
2. Advice from a *trustworthy* expert advisor can differ from advice from a *credible* expert advisor.

Credibility

- Credibility: Also combines epistemic and areatic dimensions.
 - Epistemic dimension: The credible expert must be knowledgeable in their field.
 - Aretaic dimension: The expert must be truthful.
- Motivation not characterized by goodwill, but by an orientation towards truth.

Credibility

- Trustworthy advice and credible advice can point in different directions.
 - Trustworthy advice: has to take everything that is relevant for the good of the advisee into consideration.
 - Credible advice: provides expertise in a specific field that is relevant for the decision what to do.
- Goals of trustworthy advice and credible advice differ:
 - Trustworthy advice: aims to provide potentially conclusive reasons for action.
 - Credible advice: aims to provide reasons that are relevant for practical deliberation.

Credible AI

- When turning to an AI system for advice, we do not (necessarily) want them to provide information for our own good, and we do not (necessarily) want them to take all ethical values into account that make AI system trustworthy.
- We want credible advice that provides important information in a domain that is relevant for our practical deliberation.

Credible AI – How?

- What does it mean for an LLM to be credible?
 - Epistemic dimension: the LLM must reliably provide correct information.
 - But: What does it mean for the LLM to be truthful?
 - Requires again an alternative interpretation of the areatic dimension of credibility that is close enough to our ordinary concept to make the concept applicable.

Credible AI – How?

- What does it mean for an LLM to be credible?
 - Epistemic dimension: the LLM must reliably provide correct information.
 - But: What does it mean for the LLM to be truthful?
 - Requires again an alternative interpretation of the areatic dimension of credibility that is close enough to our ordinary concept to make the concept applicable.

Credible AI – How?

- 1) No nudging: LLMs must present their advice in such a way that their role as an expert advisor is not compromised.
 - Information and advice needs to be presented in a way that does not nudge the advisee into any specific course of action.
- 2) No telling the advisee what they want to hear.
 - Requires a form of advice that not merely aligns with the advisee's preferences.
- 3) No pretending of fake expertise.
 - Fields of expertise must be clearly marked; gaps in reliable information must be explicitly addressed.

Credible AI – How?

If we have developed credible AI systems, how should we treat their expert advice?

- Larger debate in the role of expert testimony.
- Preemptionism: Expert testimony that p ought to be treated as a preemptive reason to believe p , so that the testimony replaces, rather than merely adds to, a person's epistemic reasons for or against p .
- Total Evidence Accounts: Expert testimony that p ought to be treated as a contributory reason for believing p .

Credible AI – How?

Regardless of the merits of preemptionism in the case of human experts, AI expertise seems to require a Total Evidence View.

- Only way to secure meaningful human control.
- This might provide further requirements in the design and development of LLMs that serve the specific function of advisers.

Summary

- LLMs that act as advisors should be regarded as parties in our social epistemological practices.
- As such, we need not only trustworthy AI, but also credible AI.
- This has implications for how LLMs ought to be designed – not only as reliable sources of information, but also as credible providers of epistemic reasons.



Universität Stuttgart

Vielen Dank!



Prof. Dr. Jörg Löschke

E-Mail joerg.loeschke@philo.uni-stuttgart.de
83658

philo.uni-stuttgart.de

Universität Stuttgart
Institut für Philosophie
Seidenstrasse 36 70174 Stuttgart