Trusting the machine:
Epistemic trust and anthropomorphism in generative artificial intelligence

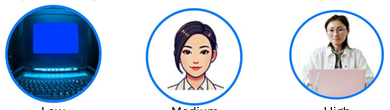
Caroline Simpson & Jonathan Fugelsang

MOTIVATION: Generative artificial intelligence (GenAI) has become **popular** for its ability to produce **fluent**, seemingly **credible content**. But how does its **perceived humanness** affect how much we **trust** what it says? We tested whether anthropomorphism affects:

- Trust in the AI **as an author**
- Trust in **the content** it produces
- And how these two trust pathways interconnect

METHOD:

Study 1: Agent vignette & avatars + blog post (n = 259)



Study 2: Interactive chatbot + blog post (n = 144)

Anthropomorphism Manipulations

- Agent description / image (Study 1)
- Chatbot emotional responsiveness (Study 2)

Measures

- METI (Author Trust; Hendriks et al., 2015)
- Message Credibility (Content Trust; Appelman & Sundar, 2016)
- Behavioural intentions novel items (Author & Content Trust)
- Godspeed Instrument (Anthropomorphism, Likeability, Competence; Bartneck et al., 2009))

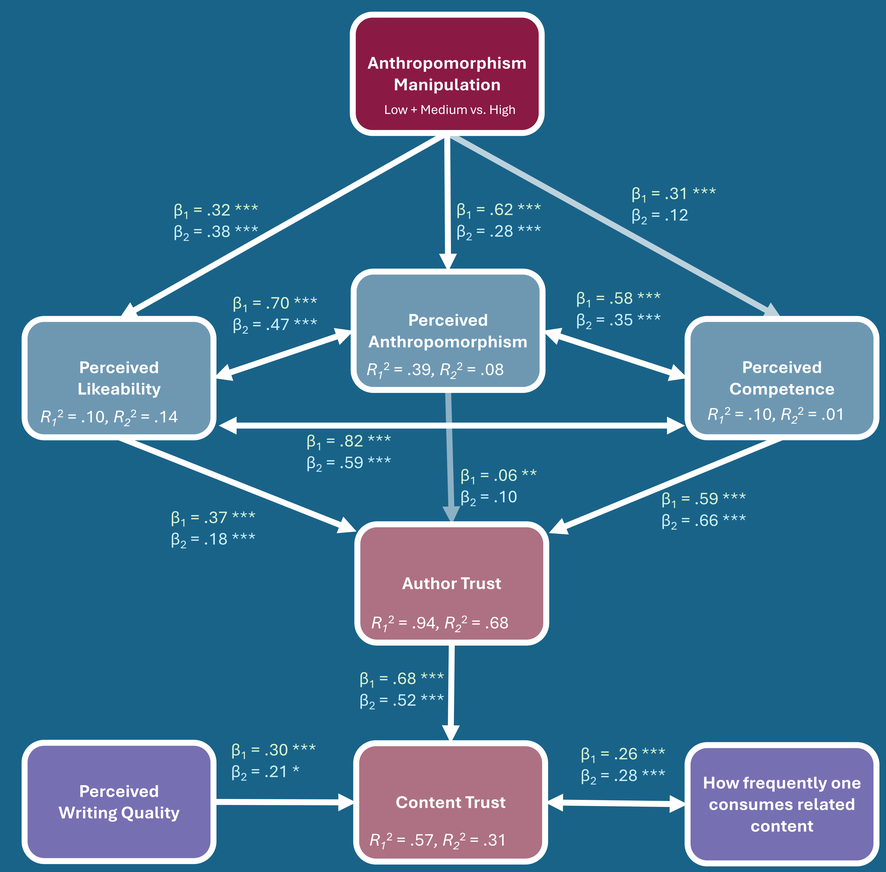
TAKEAWAYS:

- People don't necessarily trust AI content just because an AI author seems human.
- But when they trust the AI author, they trust the content more
- Anthropomorphism influences trust of the source rather than the message directly.

THE NUANCE:

- The "trust boost" of anthropomorphism is indirect
- Trust in AI content may not be easily manipulated
  - It's filtered through the perceptions of the agent
- Trust effects vary with prior content & GenAI experience

The path to TRUST ISN'T always STRAIGHTFORWARD.

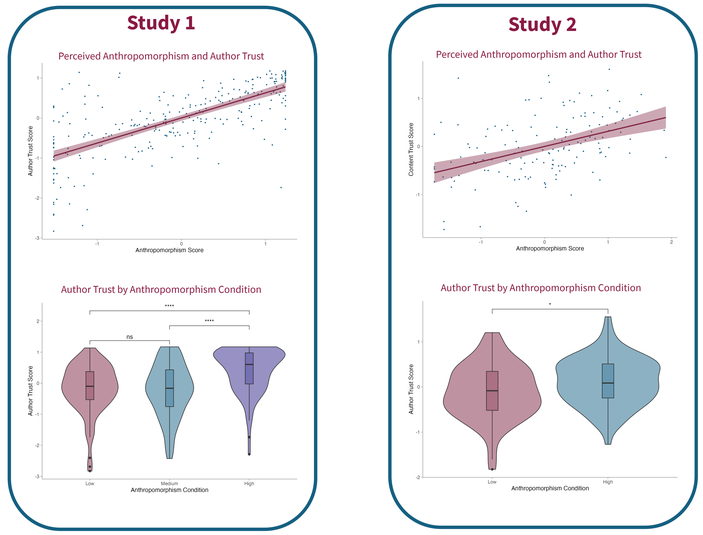


Designing GenAI for epistemic trust means shaping the user's perception of the messenger —
— not just the message.

References and more...

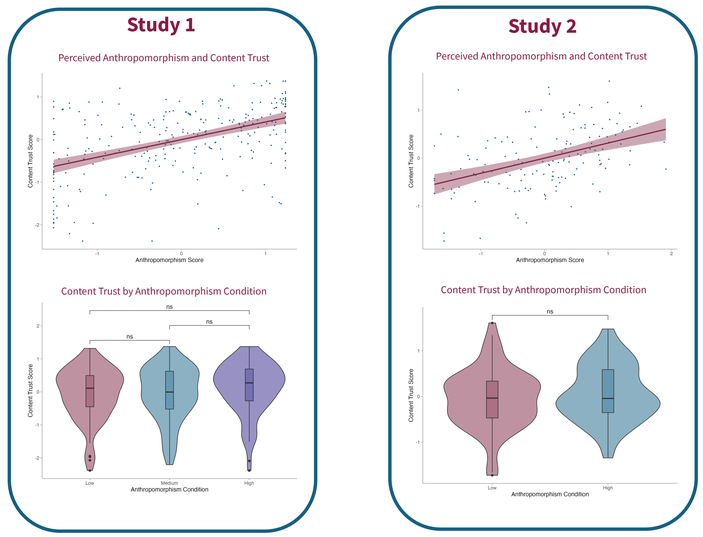
Significance: * p < .05, ** p < .01, *** p < .001

Anthropomorphism ↔ Author Trust



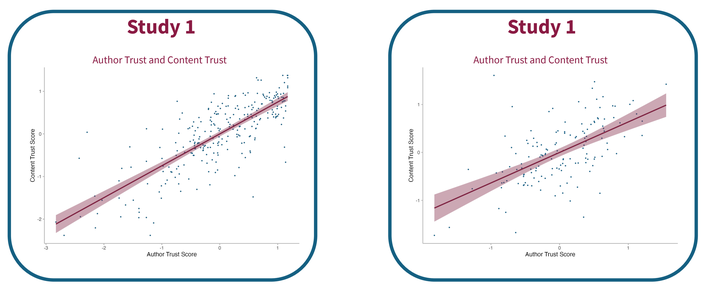
- Strong Correlation (r_1 = .70***, r_2 = .44***)
- Direct effect of manipulation in Study 2 (d = 0.40, p = .017)

Anthropomorphism ↔ Content Trust



- Moderate Correlation (r_1 = .49***, r_2 = .43***)
- No direct causal effects from manipulation

Author Trust ↔ Content Trust



- Significant positive correlation (r_1 = .78***; r_2 = .58***)

UNIVERSITY OF WATERLOO