AI as a Social Actor: Trust, Identity, and Cooperation in Human–LLM Interaction

Insights from a Repeated Prisoner's Dilemma Study

April 17, 2026

Hong Kong University of Science and Technology (Guangzhou)
Center for Metaverse and Computational Creativity (MC^2) Lab

Teaching Assistant Talk

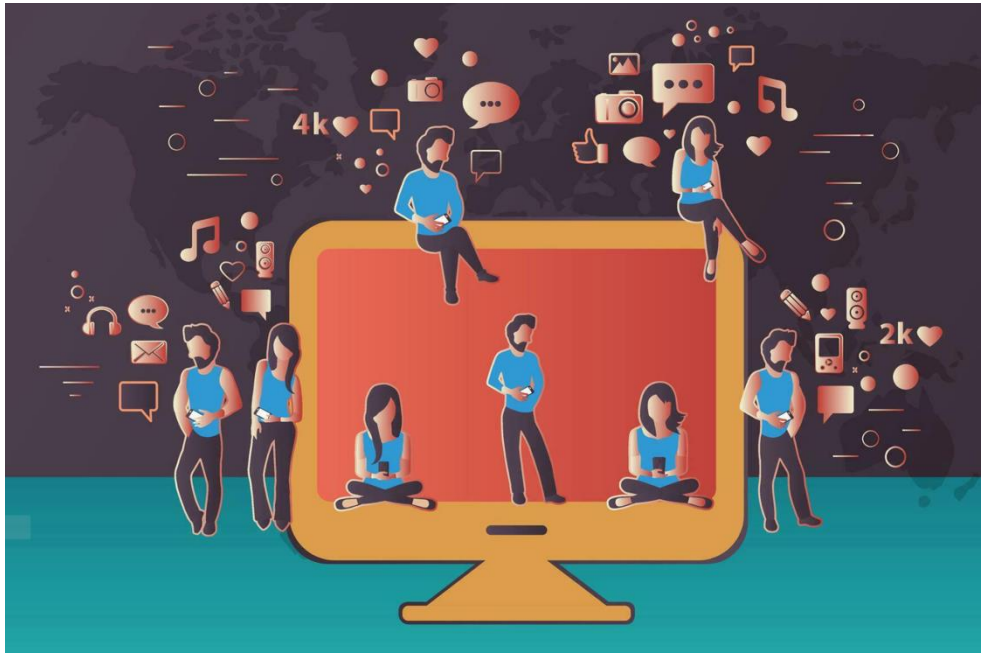
Guanxuan Jiang

Why does this matter for social media and social computing?

- AI is no longer only a tool
- AI increasingly appears as a social interface
- People interact with AI on platforms, in VR, and in collaborative settings
- The key question becomes: How do humans respond socially to AI?

What is social computing?

Social computing studies how technologies shape, mediate, and transform human social behavior.



You can think of it as:

- people interacting **through** technology
- people interacting **with** technology in social ways
- social meanings emerging around platforms, algorithms, and agents

Examples

- social media platforms
- recommender systems
- online communities
- social VR
- AI agents in collaborative environments

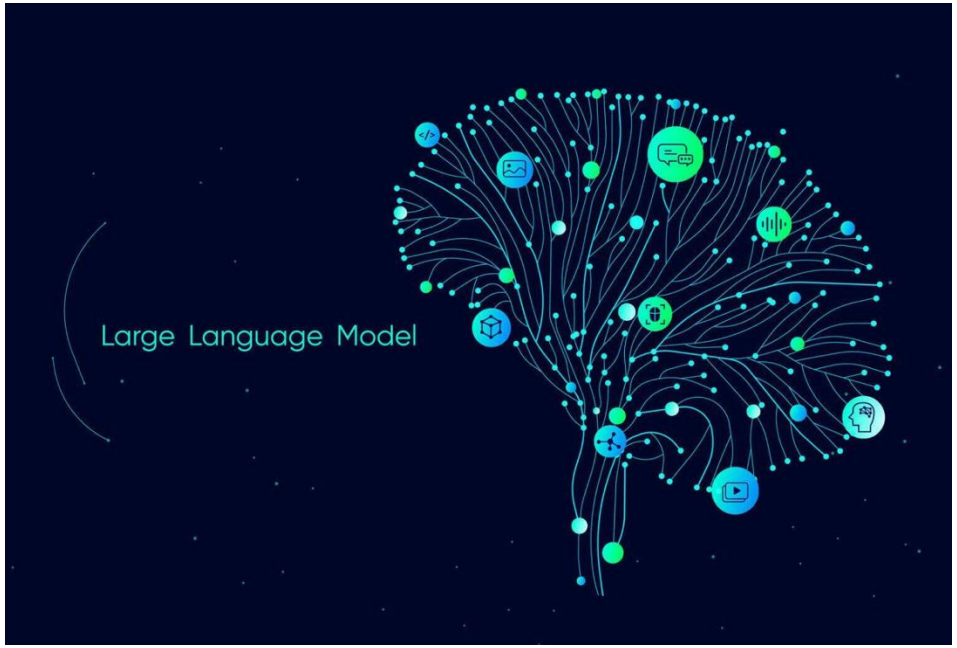
Why LLMs are different from earlier AI tools?

Earlier digital systems were often:

- rule-based
- narrow in function
- predictable
- clearly machine-like

LLM agents are increasingly:

- conversational
- adaptive
- responsive
- socially expressive
- easy to anthropomorphize

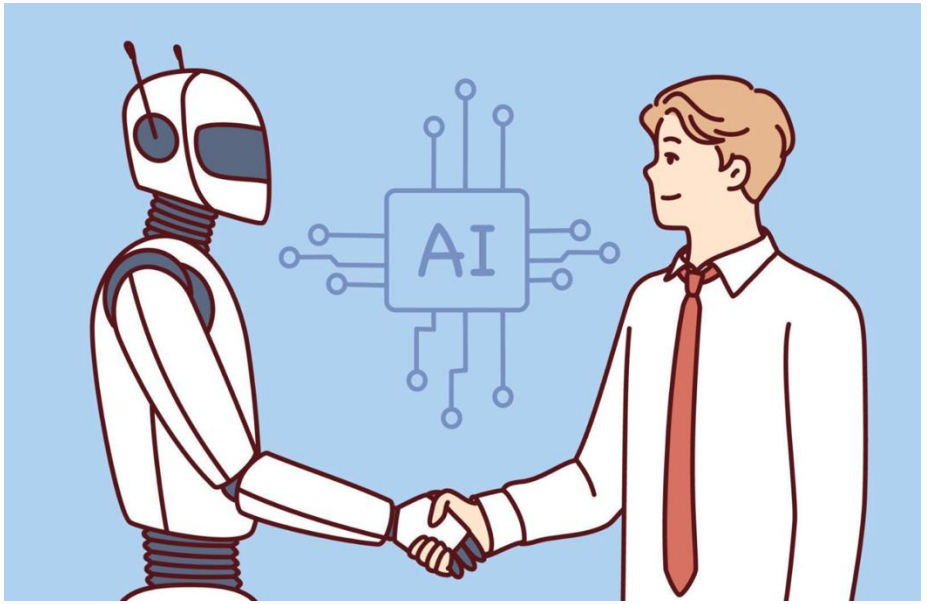


Users may respond to LLMs not only functionally, but socially.

One of the core research question

If the same AI behaves the same way, will people react differently just because of its declared identity?

- Human
- Rule-based AI
- LLM Agent



And does user gender shape these responses?

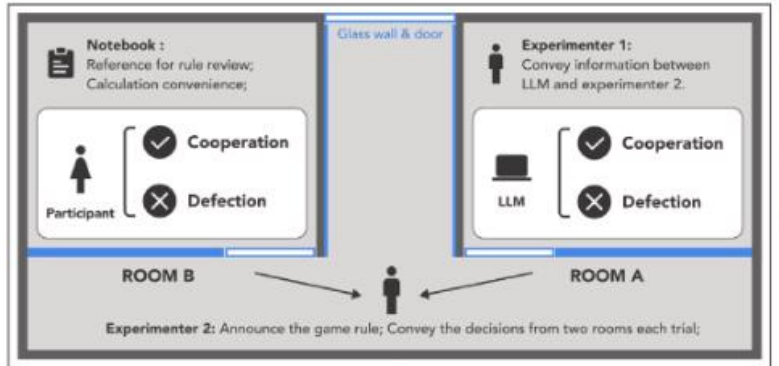
Why study mixed-motive scenarios?

- Real human–AI interaction is not always fully cooperative
- Many settings involve both alignment and tension
- We need to understand trust under uncertainty
- **Mixed-motive = shared goals + conflicting interests**
- Example paradigm: Prisoner's Dilemma

Why study mixed-motive scenarios?

Study design

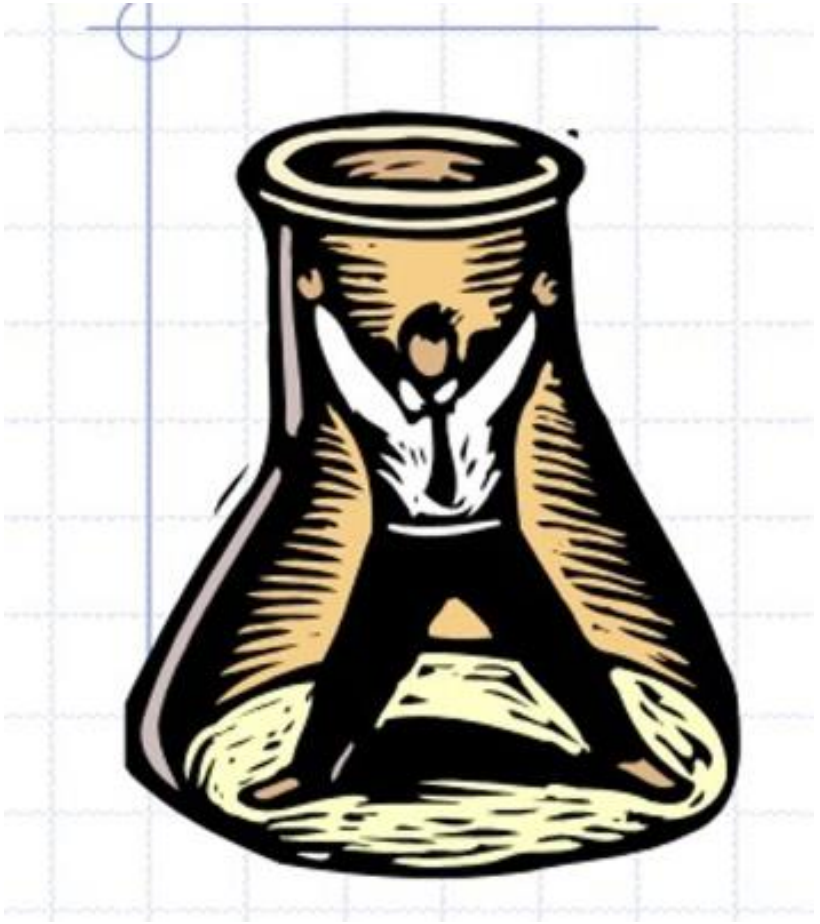
- Participants (15 Males and 15 Females)
- Three declared identities
  - Purported Human
  - Purported Rule-based AI
  - Purported LLM Agent
- Important manipulation
  - Same LLM backend across all conditions
- Task
  - Repeated Prisoner's Dilemma
  - 50 rounds × 3 sessions



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

What did we measure?

- Cooperation Rate
- Response Time
- Unsolicited Cooperation Acts
- Trust Restoration Tolerance



Key finding 1: identity matters

Finding 1: Declared identity changes cooperation

Participants cooperated more with the purported human

Declared identity significantly affected:
cooperation rate
response time
trust restoration

Identity cue alone can reshape social behavior in the prisoners' games.

Key finding 2: LLMs trigger stronger social reactions

Finding 2: LLMs evoke stronger social attribution

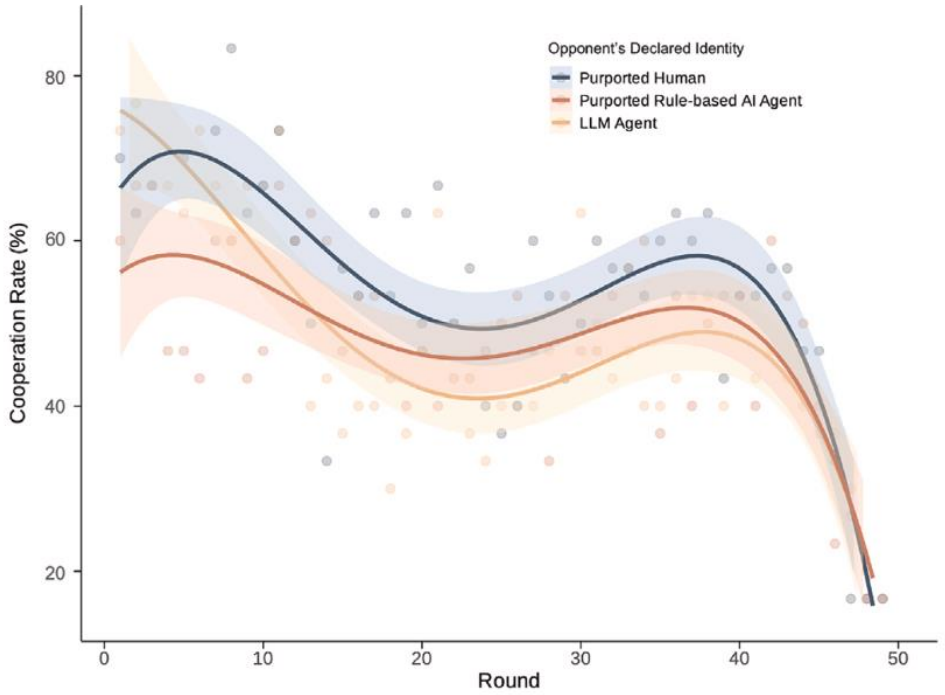
- 86.7% were more likely to use reward/punishment toward the LLM
- 83.3% expressed stronger moral condemnation toward the LLM after betrayal
- But 60.0% still viewed the rule-based AI as more reliable

LLMs are treated more like social partners, but not necessarily trusted more.

Key finding 3: higher expectations, sharper disappointment

Finding 3: Higher expectations, sharper disappointment

LLM identity → stronger initial positive expectancy → betrayal → sharper drop in cooperation



- Early cooperation toward LLMs started higher
- But declined more sharply after betrayal

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

What did the interviews reveal?

Rule-based AI

- predictable system
- logic to decode
- trial-and-error strategy
- little emotional reaction

"It's just the fixed choices of code."

LLM Agent

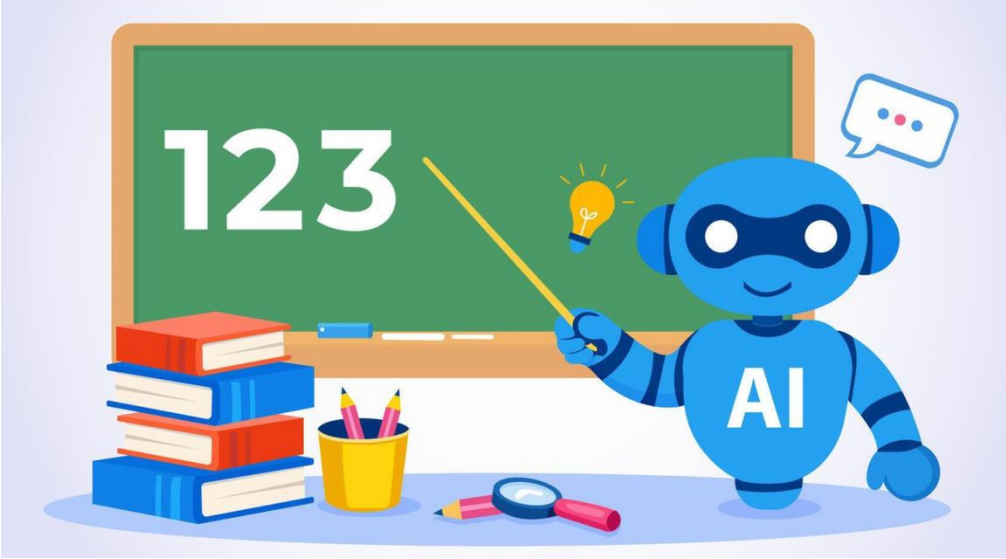
- adaptive social other
- may "learn" my behavior
- try to leave a good impression
- stronger anger after betrayal

"I want to leave a good impression and cultivate a cooperative LLM agent."

Why this matters for design

Design implications for social AI

- Identity is a design decision
- Anthropomorphism can increase engagement, but also increases risk
- Trustworthy AI needs calibration, not only capability



Relevant to AI tutors, creators' assistants, social platforms, and metaverse classrooms

Reference

[1] Jiang, Guanxuan, et al. "When trust collides: Exploring human-LLM cooperation intention through the prisoner's dilemma." *International Journal of Human-Computer Studies*(2026): 103740.

[2] Ishowo-Oloko, Fatimah, et al. "Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation." *Nature Machine Intelligence* 1.11 (2019): 517-521.

TA sessions

Thank you !