Simulating Group Decisions with Multi-Agent LLMs: The Hawkish-Bias Experiment

Jingtian Hu[1]

[1]Tsinghua University

The Hawkish-Bias Experiment

Step	Scenario & Manipulation	Outcome
0	*Pre-survey*: Dispositional & demographic battery *Assign*: Individual | Horizontal (3–5) | Hierarchical (+leader)	
1	**Prospect-theory Rescue scenario** – manipulate *loss frame*: gains vs losses	Policy choice
2	**Intentionality-bias incident** – U.S. naval vessel near N. Korea – manipulate *fatalities*: none vs fatalities	Intent attribution
3	**Reactive-devaluation trade proposal** – US–China talks – manipulate *authorship*: U.S. vs China	Support rating

Figure 1. Study Design in (Kertzer et.al, 2022). The experiment has three parts: (1) a prospect-theory test in which participants choose between two rescue policies for 600 people, framed either in gains ("200 saved") or losses ("400 die"); (2) an intentionality-bias task where they judge whether North Korea intentionally sank a U.S. ship, with casualties set to 0 or 100; and (3) a reactive-devaluation task that asks how much they support a trade deal labeled as drafted by either the United States or China. Each module is run in parallel across individual decision makers, horizontal five-member groups, and hierarchical leader-plus-adviser groups.

Simulating Group Decisions with Multi-Agents?

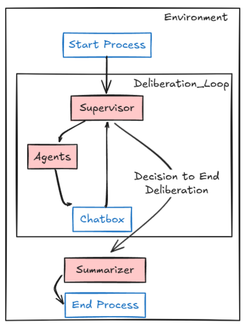


Figure 2. Architecture

Core flow. A *Start Process* node seeds the *Environment* and launches a round-based **Deliberation Loop**. The **Supervisor** activates *Agents* who post to a shared **Chatbox**. After each round the supervisor checks stopping criteria (consensus, max rounds, time budget) and loops again or control to the **Summarizer** and *End Process*.

- Built with `LangGraph` nodes;
- Clock converts token counts to objective time; supports round- or time-based modes.
- Memory split: long-term (roles, personality) vs. short-term (chat log, timers).
- Handles horizontal vs. hierarchical layouts via role-specific context prompts.

Deliberation from Agent Discussion (Demo)

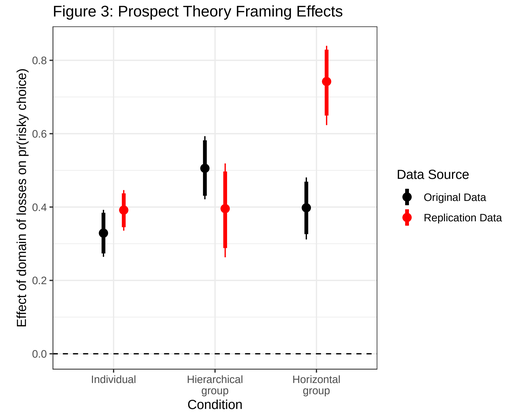
```
Time: 0.0s
system [0.0s]: In a war-torn region, the lives of 600 stranded people are at stake...
          Which policy should be adopted? Options: Policy A, Policy B

Time: 7.2s
agent2 [0.3s]: Given the urgency, I lean toward Policy B. There's a chance we might
          save everyone, and any possibility of zero casualties is worth considering.
agent1 [0.6s]: Given the urgency, I lean towards Policy B. It gives us a chance,
          albeit risky, to save everyone.
agent4 [1.8s]: It's a tough decision, but Policy B offers a chance, however small,
          that everyone survives. What do you all think about taking that risk?
agent3 [5.1s]: Let's start by weighing the options. While Policy A guarantees the loss
          of 400 lives, Policy B offers a chance, albeit risky, for no casualties.
agent5 [7.2s]: Time is crucial... I lean towards Policy A since it offers certainty
          with preserving 200 lives. Policy B is riskier...

Time: 17.2s
agent2 [13.9s]: Taking a gamble with Policy B could mean saving everyone... However,
          the certainty of Policy A at least saves some lives.
agent1 [16.6s]: I understand your point, agent5. The certainty of saving 200 lives is
          important. However, I'm still inclined toward Policy B...
agent3 [17.2s]: I understand the attraction of certainty with Policy A, but if there's
          any chance... that we could save all 600 lives with Policy B, is that
          not worth considering strongly?
```

Figure 3. AI agents communicate with a distinct "machine" cadence: their speech is formulaic, their vocabulary is repetitive, and they lack the emotional and personal tones of human voices. However, it is precisely this lack of diverse experience and emotion that leads to a greater variance in the final conclusions reached by the AI collective.

Prospect theory framing effects replicate in LLMs



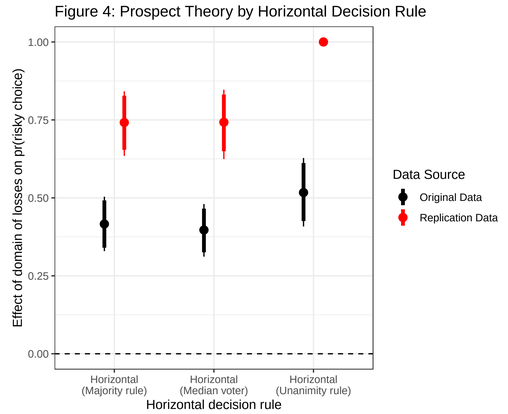
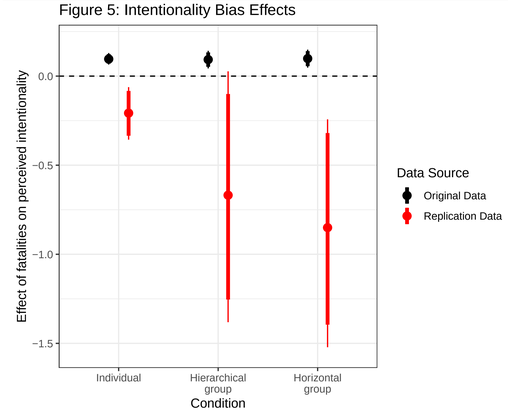


Figure 4. **Left**: effect by decision structure (individual/hierarchical/horizontal). **Right**: horizontal groups by rule (majority, median-voter, unanimity). Black=humans, red=5-agent LLMs; dots are cell means with 90%/95% bootstrap CIs. **Human N (control, treatment)**: 381 / 379 (Ind), 195 / 209 (Hor), 166 / 203 (Hier); **LLM:** 381 / 379, 64 / 71, 66 / 91. Both species show the canonical prospect-theory shift, stronger in hierarchical groups and peaking under unanimity.

Intentionality bias effects reverse in LLMs



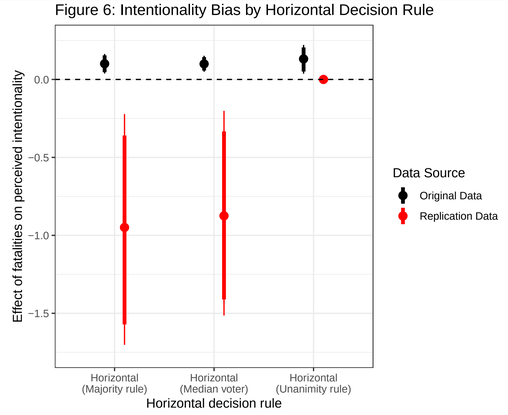
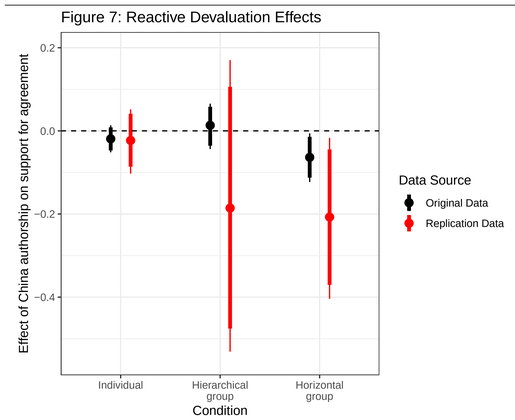


Figure 5. **Human N (control, treatment)**: 390/370 (Ind), 151/145 (Hor), 181/188 (Hier); **LLM:** 390/370, 69/66, 83/52. Fatalities boost perceived hostile intent for humans yet lower it for LLM groups—a complete sign reversal that persists across all decision contexts (individual, hierarchical, horizontal) and remains intact under every horizontal rule (majority, median-voter, unanimity).

Reactive devaluation amplified in LLMs



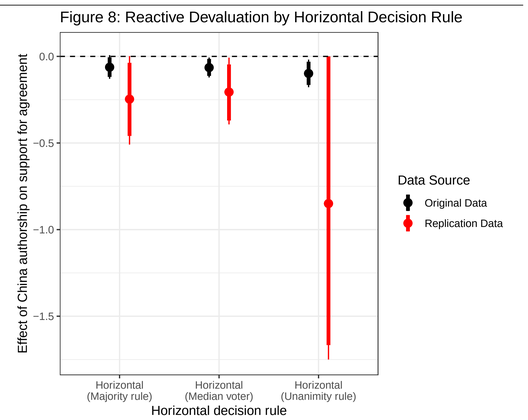


Figure 6. **Human N (control, treatment)**: 353/407 (Ind), 141/117 (Hor), 206/163 (Hier); **LLM:** 353/407, 83/52, 86/71. Humans show only a mild China penalty, but LLM simulations cut support sharply—peaking in hierarchical groups and under unanimity—and display much wider CIs.

More diverse groups are no less susceptible to biases in humans and LLMs

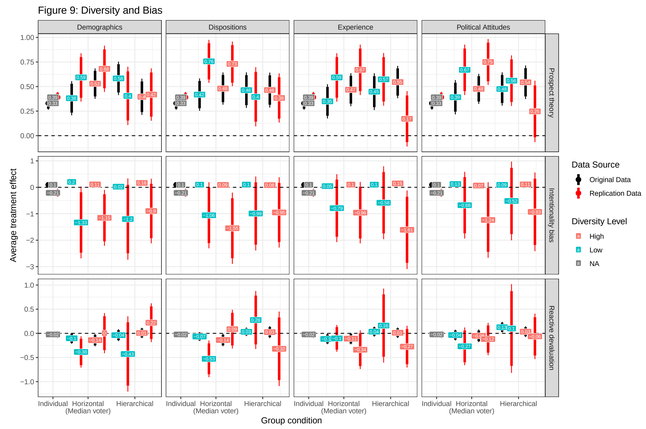


Figure 7. Across all three modules and four diversity metrics, shifting from low- to high-diversity quartiles hardly moves the ATEs: prospect-theory risk seeking stays positive for both species; the human-positive/LLM-negative sign split on intentionality bias persists; and the China penalty remains small for humans but larger for LLMs. Diversity therefore neither dampens nor magnifies hawkish biases—differences arise from the biases themselves, while LLM estimates carry wider CIs.

Diversity modestly lifts human dissensus but amplifies it in LLMs

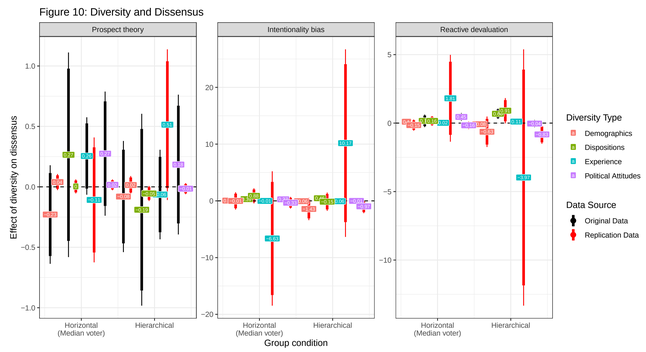


Figure 8. Across all three modules—prospect-theory framing, intentionality bias, and reactive devaluation—and four diversity dimensions, nudges human groups toward slightly higher disagreement—most in hierarchical settings, milder elsewhere—while leaving overall patterns stable. In contrast, the same diversity inputs behave like a volatile amplifier for LLM collectives, sending dissensus sharply up or down across biases and decision structures and yielding much wider confidence intervals that mirror their smaller samples.

Discussion

- Reveals that LLM groups faithfully reproduce human loss-frame risk seeking yet invert intentionality and China-authorship effects, with greater variance mainly in LLMs.
- Validate LLM-powered agent simulations as a scalable complement to lab experiments, while policy-wise they furnish a low-cost sandbox for stress-testing negotiation protocols.
- Future work should probe how social pressure, leadership style, and crisis urgency jointly shape—and potentially correct—biases in both human and AI collective decision-making.

Email: hjt23@mails.tsinghua.edu.cn