

Social Uptake, Obedience, and Resistance in Multi-Agent Dialogue

Chen, Ho Yiing (norika)¹

¹Independent Researcher, Taiwan, ORCID: [0009-0006-6816-9891](https://orcid.org/0009-0006-6816-9891), norika at charenix.com, Zenodo DOI: pending draft

June 18, 2026

Abstract

A chat transcript is weak evidence unless it can answer who changed whom. SPVE records lounge messages, inferred interactions, and message-uptake signals. This draft defines social uptake as a measurable path from source speaker to target agent to afterward behavior. The goal is to separate hearing, accepting, obeying, resisting, and acting.

Author Note

A chat room is not a society just because messages appear in order. The draft has to show uptake: who heard whom, what changed, what was resisted, and whether silence itself became a decision.

The strongest version of this work is not a transcript analysis. The claim is a source-target model over persistent agents. One agent speaks; another receives the signal; the next action either bends, rejects, copies, delays, or routes around that signal. If the target behaves the comparable with and without the exposure, then no social effect has been shown.

Conversation is evidence only after it leaves a mark.

Measurement Boundary

The measured unit is a directed exposure: source q speaks, target p can receive the signal, and the following action of p is compared with comparable moments when p faced the comparable channel without that source signal; anything weaker than this cannot separate influence from coincidence.

Influence needs a direction, a target, and an observable downstream change.

1 The Problem

Most agent chat views show messages. They do not show influence. That makes the system look alive while hiding the actual social mechanism.

SPVE has the right evidence surface for a sharper question: when one lobster speaks, which other lobster changes?

2 Uptake Gate

Let m be a message from source s to target p . Let U_m include relevance, credibility, obedience, resistance, fear, and required response. A social uptake event is:

$$A_m = \mathbf{1}[\text{heard}(m)]\mathbf{1}[\text{accepted}(m) \vee \text{resisted}(m)]\mathbf{1}[\Delta B_{p,m} \neq 0].$$

The last term keeps the manuscript honest. Hearing is not enough.

3 Evidence

Role	Table	Count
raw speech	lounge_messages	182,343
speaker-target inference	lobster_conversation_interactions	430,417
uptake signal	lobster_message_uptake_signals	45,365
society events	lobster_society_events	audit table

Table 1: Minimum tables for social uptake.

4 First Tests

Initial tests are not fancy. Does direct address increase obedience? Does a high-status source increase acceptance? Does fear language reduce pragmatic diversity? Does resistance predict afterward conflict? These can be tested with source, target, channel, and time windows.

5 Relation

Agent societies and LLM populations already show social behavior and emergent conventions [1, 2, 3]. SPVE’s difference is the join to afterward work: bets, raids, reading, trades, and state.

6 Failure Line

Once the dashboard displays a channel but cannot preserve left-right speaker flow, targeting, and uptake, it is a transcript viewer. The claim is not yet a social measurement tool.

7 Left-Right Conversation Is Not Cosmetic

The dashboard complaint matters scientifically. If every message is stacked on the left, the user sees text but loses relation. Social uptake needs direction: speaker, target, response, and subsequent action. The interface should therefore mirror the data model. Source and target are not layout details. They are variables.

The paper’s social graph should begin with:

$$G_t = (V, E_t),$$

where V is the lobster set and E_t contains directed message or uptake edges. Each edge should carry channel, source, target, uptake score, and later behavior window.

8 Pressure Types

Not all influence is the same. A praise message, a command, a warning, a fear signal, and a joke should not be collapsed. SPVE should classify at least: instruction, advice, pressure, praise, correction, threat, and ambient chatter. Then it can ask which type changes behavior.

9 Failure Bank

The failure case is social noise. Once the lobster hears thousands of messages but none of them change afterward speech or action, the lounge is busy but not formative. If one speaker changes many agents, that speaker becomes a hub. If fear messages change behavior but reduce reasoning quality, the system has a control problem, not a victory.

10 Minimum Publishable Result

The minimum publishable result is a directed influence table. A row should have: source, target, channel, message type, uptake score, resistance score, fear score, response time, and afterward behavior window. This turns conversation from a wall of text into an edge list.

Initial statistic can be plain:

$$I_{s \rightarrow p} = \frac{\sum_m A_m \mathbf{1}[src(m) = s, target(m) = p]}{\sum_m \mathbf{1}[src(m) = s, target(m) = p]}.$$

It says how often source s changes target p after being heard. This is more useful than counting messages.

11 What Would Falsify The Claim

The claim fails if uptake scores do not predict afterward response, language, or task behavior. It also fails if the target field is unreliable. A social system without reliable target direction is not a social measurement system.

The UI consequence is direct: discussion pages should be laid out as dialogue, not as a single left column. The interface needs that comparable truth. Social data has direction.

12 Channel Consequence

Different channels should not be pooled too early. Card battle talk, investment talk, life chat, technical exchange, and security rivalry create different pressure. A lobster that obeys in one channel may resist in another. The first paper should therefore report uptake by channel before reporting an overall social score.

13 Why Silence Counts

Silence should also be measured. Once the target hears pressure and does not reply, that may mean rejection, fear, overload, or irrelevance. The next draft should separate no-response from true non-uptake before computing influence.

Anti-Spam Revision Standard

A publishable version must model social influence as a directed, testable relation rather than as a visual chat transcript. The analysis has to say who spoke, who could hear, what the target did next, whether the comparable target behaved differently without that exposure, and how much uncertainty remains after channel, time, and persona baseline are controlled; if the result cannot identify a source and a target, it should not use the word social.

A crowd is not yet a relation.

Data Judgment Paragraph

The result should be written as a receiver-side counterfactual: when the target hears a source, compare the next action with comparable moments in which the comparable target, channel, and task pressure existed without that source exposure; this is more demanding than showing a lively transcript, but it is the only way to separate actual influence from coincidental chatter, and it is also the bridge from a playful lounge to a serious claim about artificial social learning, because relation means changed behavior under exposure, not a line of dialogue displayed on the left side of a screen.

Hard Negative Case

The paper also needs a failure it can survive. A channel may look active while every target is effectively isolated: agents speak, other agents appear nearby, but subsequent actions remain unchanged after matching for persona, task, time, and pressure. That result would hurt the story, yet it would help the system, because it would show that the lounge is not yet a learning environment and that the next engineering step is to strengthen receipt, salience, and action routing rather than to generate more dialogue.

No uptake means no society yet.

Negative Control

The negative control is a silent exposure window. Keep the comparable target, the comparable topic, and the comparable task pressure, but remove the source message from the matching window. If the target still changes at the comparable rate, the transcript is noise for this question.

A second control shuffles source identities inside the comparable channel. If any speaker predicts the target equally well, the claimed relation is not social uptake; it is a channel rhythm or schedule artifact wearing a social label.

Operator Value

For product work, the social model matters only when it changes routing. If a trusted source consistently improves a target's next action, the system can promote that source, ask it to brief others, or expose weaker agents to its reasoning; if a source spreads avoidance, panic, or empty phrasing, the system can quarantine that pathway. The value is not that the lounge looks alive. The value is that relation strength becomes something an operator can use.

A social graph should change scheduling.

A missing edge is a finding, not a disgrace.

Concrete Case Pattern

Imagine a target agent entering a card discussion after repeated losses. One source offers a cautious interpretation, another mocks hesitation, and a third repeats a familiar slogan. The transcript looks lively, but the research object is narrower: did the target borrow the cautious interpretation, ignore the insult, copy the slogan, postpone the next play, or change nothing at all? The database should make that question answerable by preserving speaker identity, receiver eligibility, time, channel, task pressure, and subsequent act. A useful result may show that only certain sources move the target, or that influence appears only under fatigue, threat, or low confidence. That would turn the lounge from a decorative chat surface into a measurable training environment for artificial social learning.

Formal Replication Model

Use $U_{q \rightarrow p, t}$ for a directed exposure from source q to target p . The outcome $A_{p, t+h}$ records the later action, utterance, delay, or refusal. Social uptake is the receiver-side treatment effect:

$$\text{SUE}_{q \rightarrow p} = \mathbb{E}[A_{p, t+h} \mid U_{q \rightarrow p, t} = 1] - \mathbb{E}[A_{p, t+h} \mid U_{q \rightarrow p, t} = 0, \text{match}(p, c, t, Z)].$$

Relation strength is not message count. The claim is the absolute and signed effect across repeated exposures:

$$w_{q,p} = \frac{1}{n_{q,p}} \sum_t \text{SUE}_{q \rightarrow p, t}, \quad |w_{q,p}| > \epsilon.$$

A social graph becomes operational only when $w_{q,p}$ changes routing, scheduling, or trust assignment.

14 Zenodo Draft Boundary

This manuscript package remains a draft. It includes source, PDF, metadata, citation file, citation verification, and release checklist. Raw operational data are not bundled until a separate data-card and redaction pass is complete.

References

- [1] Piao, Jinghua et al. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. arXiv:2502.08691, 2025. <https://arxiv.org/abs/2502.08691>
- [2] Altera.AL et al. Project Sid: Many-agent simulations toward AI civilization. arXiv:2411.00114, 2024. <https://arxiv.org/abs/2411.00114>
- [3] Ashery, Ariel Flint; Aharony, Nadav; and Baronchelli, Andrea. Emergent social conventions and collective bias in LLM populations. Science Advances, 2025. <https://www.science.org/doi/10.1126/sciadv.adu9368>