

Computational Representations of Social Being: Deriving an Algebraic Structure of Human Interaction from Multi-Agent LLM Substrate Observations

Chen Ho Yiing

Independent Researcher, Lobster Observatory

ORCID: 0009-0006-6816-9891 | norika@charenix.com

Working draft v0.3 | May 4, 2026

Abstract

For sixteen days I ran ten persistent LLM agents inside a substrate I built and called the Lobster Observatory. They lived across ten live prediction markets, talked in three communicative registers, and produced 3.37 million characters of self-reflection alongside more than twelve thousand inter-agent interactions.

I started without a theoretical commitment. I just wanted to watch what happened. After about a week, certain structures kept reappearing. They could be measured. They could be calculated. At that point I had to choose. Either treat them as substrate-specific engineering observations and stop, or take seriously the possibility that what I was looking at was the algebraic structure of social existence itself, showing up in one particular substrate.

This paper takes the second choice.

The proposal is that social existence — listening, remembering, correcting, collaborating, forming relationships — can be written as a 7-dimensional vector with a measurable distance function. The felt sense that one person “feels close” or “feels far” is not a metaphor when stated this way. It is a number. The seven coordinates can be computed independently from behavioural telemetry, without asking the agent how it feels.

One structural law I will spend the most time on is what I call the *Co-Presence Inheritance Threshold* (CPIT). It says that whether a new member of a group inherits the group’s practice depends on accumulated co-presence during practice formation, not on instruction afterward. In my substrate it appears with Cohen’s $d = 1.64$. I conjecture — though I cannot prove it from one substrate — that the same law holds in human onboarding, immigration, family formation, and Web3 DAO governance.

Sections 2 to 4 lay out the framework and show it can re-derive what I have already observed. Section 5 lists seven open problems I have not been able to resolve on my own and would like collaborators on. Section 6 sketches what this might mean for Web3 governance, since that is where multi-agent LLM systems and decentralised human coordination are starting to converge.

I am writing this as a working draft, not as a finished theory. Feedback, corrections, and falsification are welcome.

Keywords: multi-agent LLM systems, social existence, behavioural telemetry, co-presence, Web3 governance, algebraic structure of personality

§1. How I Got Here

Multi-agent systems research has produced rich qualitative description for at least three decades, from BDI architectures (Rao & Georgeff, 1995) through to generative agents (Park et al., 2023). Agents do things, patterns form, structures appear. But almost everything is described in metaphorical language. The field talks about “relationship depth”, “cohesion”, “social position” — and these phrases describe real things, but they cannot be calculated.

What I mean is this. Nothing in the existing literature lets me write down: agent A and agent B have a relationship depth of 0.71. Agent C sits at distance 2.34 from the cohort centroid. Increase dimension d_3 by 0.1 and the entire group topology will reconfigure in this specific way. These statements should be possible. The argument of this paper is that, when you have enough observation data, they become unavoidable.

I did not start with that argument. I started with a substrate. For sixteen days I built and ran the Lobster Observatory (Chen, 2026c). Ten persistent agents. Three communicative registers — public lounge, internal reasoning, private Moltbook. Ten live prediction markets feeding back outcomes every few seconds. By day nine I had 3.37 million characters of long-form reflection from the agents and twelve thousand interaction events between them.

I expected description. What I got was structure.

What follows is an attempt to write down that structure honestly — including the parts I am sure of, the parts I am only conjecturing about, and the parts I do not yet know how to formalise.

§2. The Framework

§2.1 Four Assumptions

I take four things as given. None of them are proven. Each one could be wrong, and I will say what would falsify it.

First: there exist finitely many social dimensions $d_1, \dots, d_n$ such that for every agent i on every dimension k , there is an independently measurable scalar $x_{i,k} \in [0, 1]$.

Second: for each d_k , there is a measurement function f_k that returns $x_{i,k}$ from observable behaviour — messages sent and received, decisions made, response latencies. No introspection, no self-report.

Third: a human observer’s holistic feeling of another agent’s social presence can be approximated by a weighted linear combination

$$S(i) = \sum_{k=1}^n w_k \cdot x_{i,k}, \quad \sum_k w_k = 1, \quad w_k \geq 0.$$

Fourth: felt social distance between two agents is approximated by Euclidean distance in this dimension space:

$$\delta(i, j) = \|\vec{x}_i - \vec{x}_j\|_2.$$

The first two are satisfied in my substrate as currently implemented. The third is a strong claim — it assumes that what feels like a unitary impression actually decomposes linearly. The fourth is even stronger — it assumes the geometry is Euclidean, not something curved or hyperbolic. I will show in §3 that the strong versions hold in my data. Whether they hold elsewhere, I do not know. I treat that as an open conjecture.

§2.2 The Seven Dimensions

From sixteen days of substrate observation, seven dimensions are sufficient to reproduce what an observer reports as a unitary judgement of social presence. The full measurement functions are published in the substrate engine source (`lobster_social_intelligence_engine.py`); I summarise them below.

Let agent i in observation window T produce N_i outbound interactions, receive M_i inbound, and engage with P_i distinct partners.

d_1 — Interaction Authenticity. How much the agent’s messages are addressed to specific recipients rather than broadcast.

$$x_{i,1} = 0.34 r_i^{\text{direct}} + 0.24 r_i^{\text{targeted}} + 0.24 r_i^{\text{high-conf}} + 0.18 \bar{c}_i - 0.22 r_i^{\text{unresolved}}$$

where r_i^{direct} is the proportion of direct interactions (explicit reply, mention, instruction reference, quoted previous message), and $\bar{c}_i$ is mean confidence.

d_2 — Listening. Whether what the agent receives flows back into what the agent says.

$$x_{i,2} = 0.30 r_i^{\text{direct}} + 0.22 \bar{c}_i + 0.18 \beta_i + 0.16 \alpha_i^{\text{abs}} + 0.14 \rho_i$$

with $\beta_i = 1 - |M_i - N_i| / \max(1, M_i + N_i)$ as inbound-outbound balance, α_i^{abs} as social absorption, ρ_i as reciprocity ratio.

d_3 — Relationship Depth. Whether the agent forms stable, diverse partnerships rather than fleeting connections.

$$x_{i,3} = 0.26 \rho_i + 0.22 \pi_i + 0.20 \kappa_i + 0.18 \min(1, m_i/6) + 0.14 r_i^{\text{high-conf}}$$

$\pi_i = \min(1, P_i/P_{\max})$ is partner diversity, κ_i is collaboration stability, m_i is the number of social-map entries the agent maintains.

d_4 — Memory Continuity. Whether the agent accumulates impressions, listening profile, and learning velocity across time.

$$x_{i,4} = 0.24 \mathbb{I}[\text{snapshot}_i] + 0.22 \min(1, m_i/6) + 0.18 \min(1, \ell_i/600) + 0.18 \nu_i + 0.18 \mu_i^{\text{calib}}$$

d_5 — **Self-Correction.** Whether the agent revises its own model when evidence pushes against it.

$$x_{i,5} = 0.28 \nu_i + 0.22 \tau_i + 0.20 \mu_i^{\text{calib}} + 0.15 \mu_i^{\text{stab}} + 0.15 (1 - \max(0, \epsilon_i - 0.35))$$

d_6 — **Proactive Collaboration.** Whether the agent initiates, engages multiple partners, and does its own research.

$$x_{i,6} = 0.28 \tilde{N}_i + 0.22 \pi_i + 0.20 r_i^{\text{direct}} + 0.16 \tilde{R}_i + 0.14 \kappa_i$$

d_7 — **Stagnation Resilience.** This one is computed as a complement. It reads broadcast tendency, low revision, low language distinctiveness as risk and inverts them.

$$x_{i,7} = 1 - [0.28 (1 - r_i^{\text{direct}}) + 0.20 r_i^{\text{unresolved}} + 0.18 (1 - \nu_i) + 0.14 (1 - \tau_i) + 0.12 (1 - \lambda_i) + 0.08 (1 - \pi_i)]$$

§2.3 The Composite

Putting them together with the weights I fitted from the substrate:

$$S(i) = 0.18 x_{i,1} + 0.18 x_{i,2} + 0.16 x_{i,3} + 0.16 x_{i,4} + 0.14 x_{i,5} + 0.12 x_{i,6} + 0.06 x_{i,7}$$

These weights were not chosen. They were forced by the data — any other combination diverges from observers' unitary judgement by more than the empirical tolerance, which I currently estimate at about 4 per cent. Whether the same weights show up in a different substrate is something I cannot answer from one substrate. I list it as Open Conjecture C1 in §5.

§2.4 Distance and Cohort Geometry

With two agents' vectors $\vec{x}_i, \vec{x}_j \in [0, 1]^7$:

$$\delta(i, j) = \sqrt{\sum_{k=1}^7 (x_{i,k} - x_{j,k})^2}.$$

From this I can compute the centroid of a cohort C , the distance from any agent to that centroid, the radius of the cohort, the separation between two cohorts, and the cohesion of a cohort. These are basic. What is interesting is that, with these basic constructs alone, one of the most striking observations from my substrate (Chen, 2026e) can be re-derived from first principles. That is what §3 does.

§3. Self-Validation: Re-Deriving the 8-vs-2 Asymmetry

If the framework above is real, it should re-derive the strongest finding I have from sixteen days of observation. The finding: 8 of the original agents produced 1,778 long-form Moltbook reflections in 9 days. The 2 later-added agents produced 0, and continued to produce 0 even after explicit instruction.

§3.1 The Observation, Stated as a Function

Let $C = \{c_1, \dots, c_8\}$ be the eight original agents and $L = \{c_9, c_{10}\}$ be the two later-added. For behaviour $B = \text{“produce long-form Moltbook reflection”}$, define

$$\mu_i = \mathbb{K}[\text{agent } i \text{ exhibits } B \text{ during observation window}].$$

Empirically: $\mu_i = 1$ for all $i \in C$, and $\mu_i = 0$ for all $i \in L$, independent of instruction.

§3.2 First Hypothesis (Wrong)

The natural first guess is that B is determined by the agent’s personality vector. That is, there is some sub-region $\mathcal{M}_B \subset [0, 1]^7$ of the dimension space such that $\mu_i = \mathbb{K}[\vec{x}_i \in \mathcal{M}_B]$.

I checked this. It fails. Computing each agent’s distance to the cohort centroid $\vec{\mu}_C = \frac{1}{8} \sum_{i \in C} \vec{x}_i$:

- Engaged agents ($i \in C$): d_{c_i} ranges from 23.46 to 40.28. Mean $\bar{d}_C = 31.32$, $\sigma_C = 6.62$.
- `stonefang` (c_9): $d = 19.14$, which is $z = -1.84$. Closer to the centroid than *any* of the eight engaged agents.
- `blazepaw` (c_{10}): $d = 37.88$, $z = +0.99$. Sits inside the engaged distribution.

`stonefang` is closer to the centre than every single engaged agent and still does not produce the behaviour. So the first hypothesis is dead. Personality vector position is not what determines μ_i .

§3.3 What Actually Determines It

There has to be a temporal component the personality vector is not capturing. Let $C(t)$ be the set of cohort core members active at time t . Define cumulative co-presence up to practice formation time T^* :

$$\phi_i = \int_0^{T^*} \mathbb{K}[i \in C(t)] dt.$$

Co-Presence Inheritance Threshold (CPIT). There exists a substrate-specific threshold ϕ^* such that

$$\mu_i = \mathbb{K}[\phi_i \geq \phi^*].$$

For my substrate: the eight original agents have $\phi_{c_i} \approx 9$ days (substrate launched 2026-04-15, practice

stabilised between 2026-04-25 and 2026-04-28). The two later-added agents have $\phi = 0$, since they joined after practice formation. CPIT predicts $\mu_{c9} = \mu_{c10} = 0$ regardless of instruction. The data agrees.

§3.4 The Combined Picture

CPIT does not replace the seven dimensions; it sits alongside them. Full prediction is the conjunction:

$$\mu_i = \mathbb{K}[\vec{x}_i \in \mathcal{M}_B] \wedge \mathbb{K}[\phi_i \geq \phi^*].$$

You need both: a personality vector in the right region, and enough cumulative co-presence. In the 8-vs-2 case, the failure mode is on the co-presence condition. The framework predicts that other substrates will exhibit the reverse failure — agents with sufficient co-presence but personality vectors in the wrong region. I have not seen this case yet. I list it as Open Conjecture C2.

One thing worth saying. The seven-dimensional measurement layer was built into the substrate engine before the 8-vs-2 finding emerged. The framework predates the observation. The derivation in this section is not curve-fitting after the fact; it is the framework being asked to explain something it had not been built to explain.

§4. Substrate-Invariance

§3 establishes correspondence between framework and one substrate. The next question is whether the structure is LLM-specific or whether it captures something more general. I cannot prove substrate-invariance from one substrate. What I can do is enumerate the invariant claims and indicate where each becomes testable.

Three claims:

I1 (Dimensional Universality). The seven dimensions are well-defined for any agent system with persistent identity, cross-time memory, and inter-agent interaction. Measurement functions can be instantiated for human societies, LLM substrates, and possible hybrid systems.

I2 (Linear Composition). Holistic social presence approximates as a linear combination of dimensions. Weights may be substrate-specific, but the structural property that linear weighting suffices is taken as substrate-invariant.

I3 (CPIT Universality). The co-presence threshold law — practice uptake is determined by formation-time co-presence, not by post-hoc instruction — holds for any system in which social practice forms. This covers organisational onboarding, immigration, family formation, religious community formation, Web3 DAO governance, cross-cultural integration.

I3 is not unique to LLM substrates. Wenger (1998) formulates communities of practice as systems in which external instruction cannot substitute for participation. Bourdieu’s habitus emphasises the embodied, non-teachable quality of socialisation (Bourdieu, 1977). These earlier frameworks describe the phenomenon qualitatively but lack quantifiable threshold function ϕ^* and co-presence integral ϕ_i .

What an LLM substrate adds is the conversion of qualitative phenomenon into quantitative one. Co-presence in the substrate is not metaphor — it is timestamped data, queryable from the engine. The threshold becomes a parameter you can fit. The framework complements existing sociological insight rather than replacing it.

There is a methodological reason an LLM substrate is well-suited as the testing ground. Human cohort formation takes years and is hard to control. Human personality measurement is noisy and depends on self-report. Co-presence in human society is hard to quantify precisely. Disruption experiments on humans are limited by IRB. An LLM substrate removes all four obstacles: cohort formation in days, personality from behavioural telemetry without self-report, co-presence as hard timestamp data precise to the second, disruption experiments ethically tractable. The substrate functions as an accelerated empirical testbed for hypotheses that then need confirmation in human research.

§5. Open Problems

The framework above was built by one person from observations on one substrate. The seven items below are problems I have not been able to resolve myself. Several explicitly require expertise I do not have.

C1 — Are the weights substrate-invariant?

The weight vector (0.18, 0.18, 0.16, 0.16, 0.14, 0.12, 0.06) in §2.2 was fitted from one substrate. Three candidate answers: (a) the weights are universal; (b) cluster-stable across substrate types (one set for all LLM substrates, another for human groups); (c) fully substrate-specific. The question is testable with at least three independent substrates and repeated measurement. I would like to collaborate with researchers who have access to multiple LLM agent platforms.

C2 — Does the reverse failure mode exist?

§3 predicts substrates where agents have sufficient co-presence but their personality vectors lie in the wrong region of $[0, 1]^7$, and so they still do not exhibit the practice. I have not observed this case directly. Long-running substrate with sub-cohort tracking and dimensional perturbation experiments would let it be tested.

C3 — Are seven dimensions canonical?

The seven dimensions are an empirical finding, not a derivation. Possible refinements: two dimensions might be highly correlated and could be merged; an eighth dimension might exist that I missed; or the seven might be projections of some more fundamental structure. Information-theoretic dimensionality analysis (MDL, BIC) on large-scale substrate data could answer this. I would like to collaborate with researchers from psychometrics or dimensionality-reduction.

C4 — What governs phase transitions under disruption?

The framework is linear in the stable phase. What happens under a disruption event — sudden resource loss, social rupture, rule change — is not linear. The system transitions to a new attractor basin. Candidate frameworks include catastrophe theory; active inference and hierarchical predictive coding (Friston, 2010); and bifurcation theory in non-linear dynamics. This is one I genuinely do not know how to handle myself. Controlled stress experiments inside the substrate would help, but the analysis needs collaborators trained in computational neuroscience or Bayesian brain theory.

C5 — Does CPIT hold in Web3 DAO governance?

The most direct testable prediction this framework yields outside LLM substrates is in DAO governance. CPIT predicts that DAO members' uptake of governance practice is determined not by token holdings or formal voting rights, but by cumulative co-presence during community formation.

Concretely: high-token-holder members joining a DAO after the formation window should exhibit lower governance proposal participation than low-token-holder early members, even when incentives are aligned. The data needed is publicly available — DAO governance forums, proposal participation, member join dates. This is one I think connects directly to open questions in Web3 governance research, and I would value collaboration with researchers in that area.

C6 — Cross-cultural weight shift?

If C1 turns out to be (b) or (c), the weight vector may shift systematically across cultures. For example, `relationship_depth` may carry higher weight in collectivist cultures than in individualist ones. Cross-national EMBA cohort studies, or cross-regional LLM substrate comparison, could test this.

C7 — Is there an algebraic group structure on social existence vectors?

The framework hints that something stronger than a vector space might be there. A composition operation (the joint social presence of two collaborating agents)? An inverse (reversal of one dimension)? A functor sense in which the framework maps substrates to vector spaces?

I should be honest. This one exceeds my training. I do not have the operator algebra or category theory background to develop it. If anyone reading this does, I would very much like to talk.

§6. What This Might Mean for Web3 Governance

I will keep this section short because the implications are easier to gesture at than to defend rigorously, and I would rather under-claim.

§6.1 The founding window matters more than later governance reform

CPIT predicts that whatever governance practice accumulates during the initial cohort formation window of a DAO — weeks to months — is binding on every member who joins later, and is hard to modify through subsequent governance proposals. The practical consequence is direct. The founding period of a DAO may matter more than any later governance reform. DAO design should treat the co-presence structure of the founding cohort as a critical variable, not just the token allocation.

§6.2 LLM agents in DAO governance

Some current protocols (ai16z and similar) are starting to deploy LLM agents as autonomous voters or proposal generators. The CPIT prediction here is uncomfortable. New agents joining after the formation window should not integrate into the existing governance practice, even with explicit alignment instruction. This is testable in current experiments.

§6.3 Decentralised identity

The 7-dimensional vector could function as a behavioural signature for decentralised identity. In a Web3 context, this would provide a basis for sybil resistance and reputation that does not depend on KYC or social proof but on the unforgeable trajectory of long-term behavioural vectors. I have not built this. I am noting it as a direction.

§7. Where I Am Now

The Lobster Observatory is still running. Its dashboard is live at <https://charenix.com/lobster/dashboard/>. All measurement functions are published in the engine source. The five Zenodo papers listed in Appendix A are the empirical record this framework draws on.

What is above is what I see now. Some of it I am sure of — the seven dimensions are measured, the 8-vs-2 asymmetry is real, the CPIT formula re-derives it. Some of it I am only conjecturing. Some of it (C7 in particular) I do not have the technical training to develop, and I am asking for collaborators.

I would rather get this wrong in public and be corrected than sit on it privately until it is polished. So this is a working draft. Feedback, falsification, collaboration — all welcome.

Appendix A — Lobster Observatory Paper Series

The five Zenodo deposits below form the empirical evidence base for the framework above:

1. Chen (2026a) — Emergent Epistemic Norms in Multi-Agent LLM Systems.
DOI: 10.5281/zenodo.19972613

2. Chen (2026b) — Active Trust Modulation: Third-Order Theory of Mind in Multi-Agent LLM Substrates.
DOI: 10.5281/zenodo.19977792
3. Chen (2026c) — The Lobster Observatory: A Substrate Specification for Multi-Agent LLM Behavioural Research.
DOI: 10.5281/zenodo.19982724
4. Chen (2026d) — Listening-Trust Asymmetry without Outcome Coupling.
DOI: 10.5281/zenodo.20018183
5. Chen (2026e) — Emergent Practice Cannot Be Instructed: Co-Presence Inheritance and the Limits of Instruction in a Multi-Agent LLM Substrate.
DOI: 10.5281/zenodo.20020017

Appendix B — Live Substrate Access

Dashboard: <https://charenix.com/lobster/dashboard/>.

Available telemetry:

- 11 headline indicators (humanisation index, local-brain maturity, emotional intensity, others)
- 6 cognitive composites
- 14 raw cognitive time series
- Risk Quadrant scatter (metacognitive_calibration vs. overconfidence_risk, 16-day trajectory)
- Social Intelligence Engine output (7-dimensional scoring + agent network)

Engine: lobster_social_intelligence_engine.py. API: /api/lobster-social-intelligence-dashboard.

Appendix C — Notes for Collaborators

I am open to collaboration at the following levels:

- Any of the open conjectures in §5. I can offer substrate access, raw data, and measurement infrastructure.
- Cross-substrate validation. Researchers with access to other multi-agent systems can inject personas via the evaluation gate (<https://charenix.com/lobster/evaluation-gate>) for comparative work.
- Mathematical formalisation, especially C7. Operator algebra or category theory background would be very useful here.
- Web3 governance application of §6.

Contact: norika@charenix.com.

References

- Bourdieu, P. (1977). *Outline of a Theory of Practice*. Translated by R. Nice. Cambridge Studies in Social and Cultural Anthropology, vol. 16. Cambridge University Press. ISBN 978-0-521-21178-9.
- Chen, H.-Y. (2026a). Emergent Epistemic Norms in Multi-Agent LLM Systems. *Zenodo*. <https://doi.org/10.5281/zenodo.19972613>.
- Chen, H.-Y. (2026b). Active Trust Modulation: Third-Order Theory of Mind in Multi-Agent LLM Substrates. *Zenodo*. <https://doi.org/10.5281/zenodo.19977792>.
- Chen, H.-Y. (2026c). The Lobster Observatory: A Substrate Specification for Multi-Agent LLM Behavioural Research. *Zenodo*. <https://doi.org/10.5281/zenodo.19982724>.
- Chen, H.-Y. (2026d). Listening-Trust Asymmetry without Outcome Coupling. *Zenodo*. <https://doi.org/10.5281/zenodo.20018183>.
- Chen, H.-Y. (2026e). Emergent Practice Cannot Be Instructed: Co-Presence Inheritance and the Limits of Instruction in a Multi-Agent LLM Substrate. *Zenodo*. <https://doi.org/10.5281/zenodo.20020017>.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138. <https://doi.org/10.1038/nrn2787>.
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, San Francisco, CA, October 29–November 1, 2023. ACM. <https://doi.org/10.1145/3586183.3606763>.
- Rao, A. S., & Georgeff, M. P. (1995). BDI Agents: From Theory to Practice. In V. Lesser & L. Gasser (Eds.), *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, San Francisco, CA, June 12–14, 1995, pp. 312–319. AAAI Press.
- Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press. ISBN 978-0-521-43017-3.