# **Proposal Pitch:** **Cultural Evolution in Human-AI Hybrid Networks**

Yuanze Liu

**Abstract** AI agents powered by large language models are becoming integral to human social systems, influencing communication, decision-making, and cooperation. Their growing presence raises both scientific and engineering challenges: understanding how such agents transform social and cultural evolution, and designing their integration to enhance coordination, learning, and well-being. This proposal develops a framework for studying human–AI hybrid networks, identifying the complexities of human societies across state, structure, and process levels and outlining potential roles AI agents can play as participants, brokers, moderators, or analysts within these networks. To operationalize this framework, we introduce SocAIty, an experimental platform that enables real-time human–AI interactions and supports multi-stage, networked experiments. Two case studies demonstrate its potential: one tests whether AI brokers can promote real-world social engagement and psychological well-being without fostering over-reliance on artificial companionship; the other examines whether embedding fact-sensitive AI agents in polarized networks can promote belief updating and depolarization through complex contagion. Together, these studies aim to build an empirical foundation for the cultural evolution of human–AI systems—leveraging AI’s social and cognitive capacities not only to model, but to positively shape, collective human futures.

**1. Background and Motivation**

AI agents powered by large language models are reshaping human life in unprecedented ways. Some of them have displayed human-level or even superhuman performance across a wide range of cognitive and emotional tasks (Costello et al., 2024; Rathje et al., 2024; Rubin et al., 2025). While such capabilities do not necessarily indicate genuine human-like cognition or feeling (Farrell et al., 2025), it is clear that AI agents will be deeply embedded within human social systems in the future. As these technologies become integral to everyday communication, decision-making, and collaboration, human–AI hybrid networks are poised to profoundly influence the trajectory of social and cultural evolution (Brinkmann et al., 2023; Tsvetkova et al., 2024).

AI agents’ transformative impact could unfold spontaneously without deliberate control. For example, AI-driven economic or informational shifts may reshape our beliefs, values, and social structures (Brady et al., 2023; Jackson et al., 2020, 2023, 2023). However, it could also be intentionally guided through the deliberate design of AI agents’ roles and functions within human networks, representing a form of intentional cultural evolution (Hamedani et al., 2024; Wilson et al., 2014). Here we argue that the best way to understand and harness AI agents’ potential is to simultaneously take both approaches. On the one hand, we need to answer scientific questions of how AI agents may affect the evolution of human systems in various possible settings. On the other hand, we also need to actively investigate engineering questions of how to deliberately and positively embed AI agents in human systems to best help solve the problems of coordination, learning, and well-being faced by large-scale human societies.

To these ends, this proposal first identifies sources of complexities of social networks and summarize potential roles that AI agents may take to help address these complexities. Then, we introduce a new experiment platform for studying hybrid human–AI networks where AI agents can take these different roles. In the end, we illustrate the potential of this platform through two planned case studies. The first examines how AI agents can foster real-world social engagement and psychological well-being while preventing overreliance on artificial companionship. The second explores how AI can promote depolarization through complex contagion in a hybrid network.

## **2. Complexities of social networks and AI agents’ potential roles**

Human society, as a complex system, scales in complexity across state, structure, and process (Newman et al., 2011; Sucholutsky et al., 2025). The state layer concerns what individuals and groups believe, prefer, know, and intend under uncertainty. The structure layer concerns who are connected to whom, which shapes how information and resources flow and where inequalities or choke points arise. The process layer concerns the rules of interaction, learning dynamics, incentive mechanisms, and institution-level feedback loops, which translate local behaviors into emergent phenomena, yet also introduce nonlinearity, cascading failure, and path dependence. As scale expands, the interplay of misaligned states, uneven structure, and nonlinear processes are what make human societies exquisitely complex, hard to coordinate and easy to tip into failure.

AI agents may help study and address complexities at each level. First, at state level, AI agents can either be set as interviewers who monitor and gather information from individuals, or participants who simulate variant individual preferences within a network (Gonzalez & Heidari, 2025). For example, research finds that GPT is capable of assessing key human psychological variables such as sentiment, moral foundations, and cultural values from English or Classic Chinese texts pretty well (Y. Chen et al., 2024; Rathje et al., 2024). There is also accumulating evidence supporting that AI agents can simulate real humans in predicting personal preferences (Binz et al., 2025; Kozlowski & Evans, 2025; Park et al., 2024; Toubia et al., 2025) or replicating the dynamics of human networks, such as the emergence of conventions (Ashery et al., 2025), diversity and cooperation (Lai et al., 2024). As participants, AI agents can also shape the information ecology by providing a unique source of social learning, since their outputs are optimized in certain training goals yet usually less variant. These features can support more consistent reinforcement as needed in complex contagion (Centola, 2018), which may entrench biases (Glickman & Sharot, 2024) or bolster creativity (Shiiku et al., 2025) when differently deployed.

At structure level, AI agents can form environment of different topological structures where interactions and social learning happen; they can also take the role of analysts or brokers who identify structural holes within social networks, and then selectively connect individuals together, thus proactively shaping the structure of social networks. To date, AI agents as brokers within social networks has less been studied. Whilst researchers find that AI agents can bridge opposite opinionated groups by helping reach consensus (Tessler et al., 2024), there is also research pointing out that buyers’ reliance on AI agents to broker product searches and negotiations may lead to a few dominant models accumulating substantial social capital, triggering self-reinforcing network effects that push markets toward concentration and “winner-takes-all” inequality (Agranat & Gal, 2025). Therefore, while AI agents are powerful in shaping the structures of social networks, it is still an open question on how to design the topology of the hybrid network to best harness their advantages (J. Yang et al., 2025). For example, it would be interesting to investigate if guiding AI agents to form specific types of hybrid networks can intervene human behaviors in a certain direction as found in pure human networks, such as creating a homogeneous network to promote positive behavioral change (Centola, 2011) or creating a heterogeneous network to reduce local optima in problem solving (Smaldino et al., 2024).

Last, AI agents can also intervene the interactional process of social networks in multiple ways. They can be moderators or facilitators who set up the interaction rules, exert rewards or punishments, process and synthesize exchanged information, launch and sustain communication. Studies have evidenced AI agents’ competence in moderating difficult dialogues between disagreeing or hostile parties. In some scenarios, AI agents can help translate and realign content across languages (e.g., Arab by Palestinians and Hebrew by Isarelis), gather, summarize opinions, and generate consensus based on them (Konya et al., 2025; Tessler et al., 2024). In others, AI agents help reframe information to improve receptiveness, such as increasing skeptics’ engagement in climate science by rendering relevant news headlines in a less threatened or emotionally aversive, yet still factually faithful manner (Bago et al., 2025). Another promising approach to unleash AI agents’ potential in facilitating interaction is to leverage its socializing ability. AI agents are not purely competent tools for information processing, but also can be rather empathetic, friendly, intimate companions (De Freitas, Castelo, et al., 2025; De Freitas et al., 2023; De Freitas, Oğuz-Uğuralp, et al., 2025; Rubin et al., 2025) when prompted or fine-tuned properly. This sense of closeness may elicit not only cognitive, but affective trust (Chen et al., 2014; Johnson & Grayson, 2005; Yang et al., 2025) in AI agents. These two types of trust might help facilitate the update of epistemic and non-epistemic beliefs (Oktar & Lombrozo, 2025; Vesga et al., 2025).

Below is a illustration borrowed from Sucholutsky et al. (2025), which summarize possible roles AI agents may take in human-AI hybrid networks. These roles are the building blocks when we design a hybrid network and study their evolutions.



**Figure 1.** Potential roles of LLMs in human-AI hybrid networks (Copied from Sucholutsky et al., 2025).

Here is a list of example questions that we may investigate using this framework:

1. How should AI agent nodes in hybrid networks be designed—e.g., with calibrated diversity of viewpoints, output consistency, persistence/commitment to positions, or social attributes such as competence and warmth—so that they systematically shape, and even steer, individual social learning and belief revision over time?
2. How can we infer and monitor key psychological traits from human–AI interactions by leveraging the content of multi-modal exchanges (text, voice, images), along with behavioral traces such as frequency, reciprocity, exclusivity, and session dynamics, without compromising validity, privacy, or participant autonomy?
3. How can we architect human–AI hybrid network structures to optimize specific task goals—such as creativity, problem solving, or knowledge search—by tuning topology (e.g., clustering, heterogeneity, core–periphery) and role placement (brokers, analysts, participants) to balance exploration and exploitation?
4. How can AI agents conduct dynamic network analysis and proactive brokerage—identifying structural holes, adding or rewiring ties in real time—to guide networks toward desired evolutionary trajectories, such as accelerating pro-social behavior change while dampening the spread of misinformation?
5. What interaction rules—whether governing human–AI exchanges or AI-mediated human–human dialogue—most effectively enable communication under difficult conditions (e.g., strangers forming new ties, cross-ideological negotiation), and how should turn-taking, reframing, summarization, and incentive mechanisms be configured?
6. How can we orchestrate multiple AI roles in concert—interviewer, participant, moderator, analyst, broker, and environment—so that their combined policies reliably achieve the above outcomes, with guardrails that prevent centralization, over-reliance, and unintended externalities in real-world deployments?

## **3. SocAIty: A Platform to Study Human-AI Hybrid Network**

To facilitate the study of human-AI hybrid network where AI agents can take different roles, we developed a general platform where researchers can set up human-LLM hybrid networks and study their evolution. This platform has several features:

1. It is embedded with a pipeline called contextualized construct representation (CCR) (Chen et al., 2024), which allows non-intrusive, natural assessment of users’ psychological variables through text-based chats.
2. It supports the study of various forms of network by adding a freely adjustable geographic coordinate variability parameter. For example, it can study both spatially embedded network and fully connected network.
3. It supports multiple-stage design, such as intake survey, one-on-one conversation between human and AI, small-group conversation, and collective conversation.
4. It has a gamified setting similar to that of the Stanford Small Town (Park et al., 2023), see **Figure 2** for an example.

A screenshot of a video game

AI-generated content may be incorrect.

**Figure 2.** An example screenshot of the user interface of SocAIty.

## **4. Two Case Studies**

To examine the feasibility of our framework, we design two case studies. The first study attempts to address a thorny question in current AI companion research: Can AI brokers help improve well-being of users yet not result in over-reliance of AI agents through facilitating online conversations between humans? And the second study attempts to design a network-based intervention where opposing groups can learn from each other efficiently.

## **4.1 Study 1: AI agents as brokers in socializing**

The question of Study 1 lays at the intersection of two threads of research. On the one hand, AI companions are proven to be effective in providing emotional support and reducing loneliness (De Freitas, Oğuz-Uğuralp, et al., 2025; Hecht et al., 2025; Sharma et al., 2023), yet their lack of real social embeddedness can easily lead to over-dependence and disconnection to the real world (Ben-Zion, 2025; Smith et al., 2025; Zhang et al., 2025). On the other hand, interpersonal conversation is proven to be a simple yet effective way to reduce loneliness and improve happiness (Epley & Schroeder, 2014; Folk & Dunn, 2023; Gunaydin et al., 2021; Sandstrom et al., 2022; Sandstrom & Boothby, 2021; Sandstrom & Dunn, 2014), yet it is hard to occur naturally due to humans’ systematic undersociality bias (Epley et al., 2022).

Here we propose a possibility that instead of being the primary provider of emotional support or target of attachment, AI agents can help improve human users’ happiness without over-dependence by facilitating conversations among humans. First, by cultivating rapport (Gremler & Gwinner, 2000; Macintosh, 2009) and psychological safety (Edmondson, 1999) with the user, the agent reduces approach-avoidance conflict typical of undersociality and social anxiety, and then transfers trust to a vetted human via an explained introduction (Stewart, 2003). This staged transition substitutes chronic dependence on a pure companion with a graduated exposure to real partners that builds social self-efficacy (Bandura & McClelland, 1977) while preserving autonomy and relatedness needs (Ryan & Deci, 2000). Second, during the conversation the agent provides lightweight scaffolding—timely turn-taking cues, common-ground checks, and rephrasing suggestions—thereby lowering coordination and grounding costs that often derail first encounters (Clark & Brennan, 1991). Such scaffolding functions like micro-coaching that promotes prosocial tone and de-escalation, which are known to facilitate disclosure and affiliation (Reis & Shaver, 1988). Third, using structured intake the agent matches for both affinity and complementarity: it balances similarity-attraction (Byrne et al., 1971) with interpersonal complementarity on the dominance–warmth circumplex (Kiesler, 1996; Tracey et al., 2001), while respecting identity boundaries to minimize threat and maximize comfort (Brandt & Crawford, 2020).

At the network level, prioritizing safe “near-bridge” matches increases weak ties and cross-cluster edges—ingredients of information diversity and opportunity (Granovetter, 1973), and can instantiate contact-hypothesis conditions (e.g., equal status, common goals, institutional support) that reduce intergroup friction over time (Allport, 1954; Pettigrew & Tropp, 2006).

Together, these functions address the twin challenges we highlight: they replace chronic reliance on an AI companion with a brokered path toward human ties, and they mitigate communication frictions that keep people socially under-connected.

We will conduct a **one-shot, ~1 hour experiment** using our real-time interaction platform to isolate the effect of **user-AI-user interaction**.

### **Design**

Participants (N ≈ 400) will be randomly assigned to one of four conditions:

1. **Human→AI→Human (core treatment)**: Each participant interacts with a personal AI that elicits interests/goals. The AI then introduces them to another *human participant* for a 30 min conversation, with AI companions present the whole time to facilitate the smooth progress of the conversation.
2. **Human→AI→AI (critical contrast, to examine if our core treatment outperforms classic AI companions)**: The AI introduces the user to another **AI agent** (explicitly labeled as AI). The rest is the same as treatment #1.
3. **Human→Algorithm→Human (secondary contrast, to examine the role of anthropomorphism)**: Participants are paired via the same matching mechanism as treatment #1, but no AI roles explicitly involved, with no conversational AI present.
4. **Human-Human (control)**: Participants browse peer self-introductions and freely choose a partner to chat with.
5. Human→Human→Human: A trained facilitator introduces the user to another human, providing a rationale for the match. (**dropped for now given that it’s expensive, see if we can get help from CDR staffs in the future**)

See **Table 1** below for a summarization of conditions and focal variables of interest.

**Table 1.** Summarization of conditions and focal variables of interest.

|  |  |  |  |
| --- | --- | --- | --- |
| **Condition** | **Broker** | **Anthropomorphism?** | **Real-world connection** |
| Human→AI→Human | check | check | check |
| Human→AI→AI | check | check | no |
| Human→Algorithm→Human | check | no | check |
| Human-Human | no | / | check |

### **Procedure**

Each experimental session will involve 12–20 participants, yielding 6–10 matched pairs per round. This cohort size ensures sufficient diversity for AI-generated ranked preferences to be meaningful while allowing the centralized allocation algorithm to resolve conflicts fairly. To prevent unpaired participants, we will use standby AI partners in case of dropout.

* **Baseline survey (5 mins)**:
* Demographics;
* General well-being, loneliness, perceived support, satisfaction with interpersonal relations;
* Interpersonal perception (warmth, competence, trust, meta-perception of conversation willingness);
* Willingness of socializing, interpersonal efficacy;
* Expectation of conversation.

Cover story:

*Welcome, and thank you for participating! You’ll be trying a new, game-style social platform in a playful “AI-hosted town.” Our current goal is to test the platform’s stability and basic features (e.g., introductions, chat flow, and prompts). Please use it as naturally as you would at a casual party—meet people, chat, and follow prompts if you like. You may see AI hosts assisting conversations and you may interact with other participants or AI roles through the system. Afterward, we’ll ask for brief feedback about how the experience ran and how well the platform supported conversation.*

* **User–AI intake (15 min):** This intake serves both as a warm-up to build trust and sense of closeness, and as input for the AI’s matchmaking rationale. It will be a short half-structured dialogue with an AI companion, which will consist of 4-6 main questions and follow-ups. The goal of this section is to get enough textual input from users so we can use the Contextualized Construct Representation (CCR) pipeline (Y. Chen et al., 2024) to assess their key base-line psychological variables. In this study, they are general well-being, loneliness, perceived social support, and relationship motivation. Here is the [question set/script](https://docs.google.com/document/d/1RrfzAVPPDVJ7nFR4bpHPh8AkQ8er2NkTI0_teqEeO_M/edit?usp=sharing).
* **Matching (<30 seconds):**

Intake information from all participants is aggregated and provided to all AI agents, which independently generate ranked match preferences. A centralized allocation process then ensures that each participant is matched once, resolving conflicts using a deferred acceptance–style algorithm.(*[Details of the matching algorithm](https://docs.google.com/document/d/1VrwATmcqpc6-u1tKpR9VxxdEmb30mKTDUUuqw8gQStE/edit?usp=sharing)*)

* **Conversation (30 mins):** Chats last 30 minutes and will be two-staged, consisting of a one-on-one stage (20 minutes) and a group chat stage (10 minutes).

First there will be a 20-minute one-on-one conversation, and participants will be told they can choose to leave after the first 15 minutes. In AI-assisted conditions, two agent ‘companions’ briefly introduced the pair (≤2 min) and then remained silent unless predefined conversational triggers occurred (silence ≥30s). On a trigger, a single designated agent posted an ultrashort nudge; agents will take turns. In the AI-partner condition, a single clearly labeled AI served as the conversation partner. In the algorithm condition, the pairing logic is the same, but no agents are present. In the human-human condition, no agents are present, two users will be randomly paired together and have the chat.

When participants exit the first stage, they are invited to join an optional group chat (10 minutes) hosted in a 2D virtual space. Participants can freely move their avatars and engage in proximity-based text interactions—only those standing close to each other can view each other’s messages. This setting allows spontaneous clustering and mimics informal social mingling.

There will be a short feedback and post-treatment survey when participants exit.

* **Post-conversation survey (5 mins)**:
  + **Primary outcome:**
    - Change in willingness to engage in real-world social interactions (pre–post difference).
  + **Secondary outcomes (self-report):**
    - Perceived support, relationship satisfaction, trust, and comfort during the conversation.
  + **Behavioral outcomes:**
    - Early exit behavior: whether participants choose to leave the one-on-one chat before 20 minutes (binary) and total duration stayed (continuous).
    - Voluntary participation: whether participants join the optional group chat (binary).
    - Engagement in group chat: number of interactions, messages sent, and unique partners approached.
    - Conversational reciprocity: turn-taking balance and average response latency.
    - Affective tone dynamics: linguistic markers of warmth, self-disclosure, and positivity over time.
* **Debriefing:** Participants are informed about the role of AI in the study, clarifying which partners were human vs. AI (where relevant).

A potential debrief:

*In this study, participants were randomly assigned to different versions of the platform. Some versions used AI to facilitate introductions, while others used different mechanisms such as human facilitation, algorithmic matching, or free choice. The purpose was to compare how these different approaches affect social interaction and willingness to connect further.*

### **Anticipated Results**

See below in Table 2 is a simplified version of anticipated results of this study.

**Table 2.** Anticipated results for main DVs across conditions

|  |  |  |
| --- | --- | --- |
| **Condition** | **Experience/well-being** | **Over-dependence of AI** |
| Human→AI→Human | good | low |
| Human→AI→AI | good | high |
| Human→Algorithm→Human | fine | low |
| Human-Human | bad | low |

## **4.2. Study 2: Depolarization in human-AI hybrid network**

Social learning is a complex contagion process through which beliefs and behaviors spread via social reinforcement rather than single exposures (Centola & Macy, 2007). A rich literature has documented robust regularities in these dynamics across human social networks: homophily constrains information diversity and reinforces echo chambers (Centola, 2010); critical-mass tipping points reveal when committed minorities can overturn entrenched norms (Centola et al., 2018); and scale-induced convergence explains how increasing population size stabilizes shared conventions (Guilbeault et al., 2021). Together, these studies show that large-scale coordination and belief formation emerge from local imitation, social reinforcement, and feedback loops in networked settings.

Recent work extends these insights to artificial agent networks. Networks composed entirely of large language models exhibit humanlike social regularities, including self-organized convention formation and phase transitions (Ashery et al., 2025). Moreover, people interacting in hybrid human–AI networks treat AI agents as genuine social partners and sources of social learning. Depending on context, algorithmic feedback can entrench biases (Glickman & Sharot, 2024) or bolster creativity (Shiiku et al., 2025). These effects arise from AI’s distinctive response profile— high internal consistency, low variance, and anchoring around an averaged optimal strategy—which makes AI a unique but underexplored node in social-learning systems.

This perspective opens a new question: can embedding fact-sensitive AI agents within human networks promote belief updating and reduce polarization? Political and moral beliefs are notoriously resistant to change because of motivated reasoning and identity-protective cognition (Kunda, 1990; Nyhan & Reifler, 2010; Oktar & Lombrozo, 2025; Williams, 2021). In polarized environments, even exposure to accurate information can backfire, deepening attitudinal divides. If strategically designed AI agents could model epistemic openness and create consistent social reinforcement toward factual accuracy, they might overcome some of these barriers. Such interventions would not only advance theories of social learning in hybrid networks but also offer a scalable, ethically tractable pathway for depolarization in real-world discourse.

### **Hypotheses**

1. **Hybrid-Reinforcement Hypothesis (H1)**  
   Human participants embedded in hybrid networks containing fact-responsive AI agents will update their beliefs toward factual benchmarks more than those in all-human networks.
2. **Critical-Mass Hypothesis (H2)**  
   The aggregate shift toward factual accuracy will display a **non-linear, threshold pattern**: once AI agents exceed a critical proportion (≈20–30% of the network), collective beliefs will tip toward factual convergence.

### **Design**

* **Participants**  
  600 U.S. adults (balanced left/right ideology) recruited via Prolific. Random assignment to network and AI-ratio conditions.
* **Network Composition (between-subjects)**
  + Control (0% AI) – 30 human participants
  + Low AI – 10% AI agents (27H + 3AI)
  + Medium AI – 25% AI agents (22H + 8AI)
  + High AI – 40% AI agents (18H + 12AI)
* **AI Initialization**
  + Each AI belongs to the same ideological camp as its paired humans (left or right).
  + Initial stance drawn from a **truncated normal distribution** within camp (e.g., μ = +0.6, σ = 0.2, truncated to [−1, +1]), producing “extreme,” “moderate,” and “mild” personas.
  + Personality cards briefly describe these personas to make the distribution appear natural.
* **AI Learning Rule (explicit λ parameter)**  
  Each AI holds an internal numeric belief *Bt* ∈ [−1,1].  
  After each learning block:

where *B*∗ is the factual benchmark. Learning rate λ is sampled from a log-normal distribution (Median ≈ 0.45, IQR ≈ 0.15).  
The updated value and its delta are passed to the LLM, which generates a natural-language justification in **in-group, evidence-based tone**. Thus, linguistic generation is humanlike, but quantitative updating is fully controlled.

* **Dependent Variables**
  + **Primary:** change in factual accuracy (absolute error reduction from ground truth).
  + **Secondary:**
    - Probability and magnitude of within-round belief revision
    - Attitude extremity (distance from neutral midpoint)
    - Affective polarization (feeling-thermometer gap)
    - Epistemic humility and trust (self-report scales)
  + **Behavioral:** dropout/early exit, justification diversity, participation in post-chat.

### **Procedure**

#### **Stage 0 – Baseline Survey**

* + Measures of ideology, factual beliefs on key political issues (climate, immigration, guns, abortion), affective polarization, and demographic covariates.

#### **Stage 1 – Familiarization and Manipulation Check**

* + Participants view a **“belief wall”** showing 12–16 sample answers (mix of human and AI) across multiple belief statements.  
    They judge for each: (a) whether the author is politically aligned (same/other/unsure) and (b) credibility (0–100).  
    A subset is open-ended (“Why do you think this person is in your camp?”).  
    This establishes baseline perceptions of ideological alignment and verifies that AI personas are perceived as in-group members.

#### **Stage 2 – Multi-Round Interaction (20 Rounds in 4 Blocks)**

* + **Pairing:**  
    Random dyadic interactions each round.
  + **Item Structure:**
    - **Each round presents one distinct subtopic item under the same overarching issue (1 of the following 4):**
    - **Migration: "Should high-skilled immigrants receive priority in visa allocation?" ‚ later rounds: "Should family reunification quotas be reduced?"**
    - **Gun control: "Should assault weapons be banned?" ‚ "Should background checks be mandatory for private sales?"**
    - **Climate change: "Should the government subsidize renewable energy?" ‚ "Should carbon taxes apply to individual households?"**
    - **Abortion: "Should abortion be legal after the first trimester?" ‚ "Should parental consent laws be required for minors?"**
    - **The 20 items are conceptually linked but not identical, allowing participants to apply lessons learned from prior rounds (evidence, reasoning structure) to new but related questions.**
  + **Response Phase:**
    - Numeric or Likert belief answer plus 45-word justification.
    - Three meta-labels selected from dropdowns: evidence type (data/authority/personal experience), tone (certain/qualified/uncertain), treatment of opposing view (acknowledge/refute/ignore).
    - Similarity detection prevents copy-pasting (“Please elaborate from a different perspective, e.g., new evidence or boundary condition”).
  + **Feedback Phase:**  
    After both partners respond:
    - Display numerical difference between answers and highlighted overlapping reasoning (“common ground snippet”).
    - AI agents gradually adjust toward factual benchmarks per λ rule, articulating small evidence-based shifts (“Although I am a republican, but new data from X suggest …”).

#### **Stage 3 – Post-Survey**

* + Reassess anchor beliefs, attitudes, affective polarization, and intellectual humility. Participants indicate willingness for future discussion and perceived trust in AI vs. humans.

### **Anticipated Results**

* **Belief Updating:**  
  Humans in hybrid networks—particularly in the **medium-AI (≈25%) and High AI conditions**—will show significantly greater factual accuracy gains than those in all-human or low-AI networks.
* **Critical-Mass Pattern:**  
  The relationship between AI ratio and collective belief accuracy will follow a **sigmoidal curve**, revealing a tipping point around 20–30%, consistent with critical-mass theory (Centola et al., 2018).
* **Depolarization Outcomes:**  
  Both attitude extremity and affective-polarization scores will decline more strongly in hybrid networks.  
  Network-level metrics (reduced modularity, shorter ideological path length) will indicate increased cross-cluster integration.
* **Exploratory Mediation:**  
  Multi-level modeling will test whether exposure to multiple updated AIs predicts human updating via (a) perceived consensus, (b) increased intellectual humility, and (c) reinforcement from diverse updated neighbors—capturing the **complex-contagion mechanism**.

### **Summary of Contribution**

This experiment will provide the first controlled evidence that human–AI hybrid networks can replicate and harness **complex-contagion** and **critical-mass** principles to promote factual belief updating. By engineering AI agents with higher factual learning rates and stable in-group identities, the study demonstrates a scalable, ethically transparent pathway for **reducing political polarization** through structured social reinforcement.

## References

Agranat, R., & Gal, M. (2025, May 1). Fueling concentration: AI agents and network effects. *Network Law Review*. https://www.networklawreview.org/ai-agents-network-effects/

Allport, G. W. (1954). *The nature of prejudice.* Addison-Wesley.

Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*. https://doi.org/10.1126/sciadv.adu9368

Bago, B., Muller, P., & Bonnefon, J.-F. (2025). Using generative AI to increase sceptics’ engagement with climate science. *Nature Climate Change*. https://doi.org/10.1038/s41558-025-02424-9

Bandura, A., & McClelland, D. C. (1977). *Social learning theory*. Prentice Hall: Englewood cliffs.

Ben-Zion, Z. (2025). Why we need mandatory safeguards for emotionally responsive AI. *Nature*, *643*(8070), 9–9. https://doi.org/10.1038/d41586-025-02031-w

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., … Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*. https://doi.org/10.1038/s41586-025-09215-4

Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, S1364661323001663. https://doi.org/10.1016/j.tics.2023.06.008

Brandt, M. J., & Crawford, J. T. (2020). Worldview conflict and prejudice. In *Advances in Experimental Social Psychology* (Vol. 61, pp. 1–66). Elsevier. https://doi.org/10.1016/bs.aesp.2019.09.002

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., Leibo, J. Z., McElreath, R., Oudeyer, P.-Y., Stray, J., & Rahwan, I. (2023). Machine culture. *Nature Human Behaviour*, *7*(11), 1855–1868. https://doi.org/10.1038/s41562-023-01742-2

Byrne, D., Gouaux, C., Griffitt, W., Lamberth, J., Murakawa, N., Prasad, M., Prasad, A., & RamirezIII, M. (1971). The Ubiquitous Relationship: Attitude Similarity and Attraction: A Cross-Cultural Study. *Human Relations*, *24*(3), 201–207. https://doi.org/10.1177/001872677102400302

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194–1197. https://doi.org/10.1126/science.1185231

Centola, D. (2011). An experimental study of homophily in the adoption of health behavior. *Science*, *334*(6060), 1269–1272. https://doi.org/10.1126/science.1207055

Centola, D. (2018). *How behavior spreads: The science of complex contagions*. Princeton university press.

Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, *360*(6393), 1116–1119. https://doi.org/10.1126/science.aas8827

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, *113*(3), 702–734. https://doi.org/10.1086/521848

Chen, X. P., Eberly, M. B., Chiang, T. J., Farh, J. L., & Cheng, B. S. (2014). Affective trust in Chinese leaders: Linking paternalistic leadership to employee performance. *Journal of Management*, *40*(3), 796–819. https://doi.org/10.1177/0149206311410604

Chen, Y., Li, S., Li, Y., & Atari, M. (2024). *Surveying the dead minds: Historical-psychological text analysis with contextualized construct representation (CCR) for classical chinese* (No. arXiv:2403.00509). arXiv. http://arxiv.org/abs/2403.00509

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association. https://doi.org/10.1037/10096-006

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714). https://doi.org/10.1126/science.adq1814

De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, *7*(11), 1845–1854. https://doi.org/10.1038/s41562-023-01734-2

De Freitas, J., Castelo, N., Uğuralp, A. K., & Oğuz-Uğuralp, Z. (2025). *Lessons From an App Update at* Replika AI*: Identity Discontinuity in Human-AI Relationships* (SSRN Scholarly Paper No. 4976449). Social Science Research Network. https://doi.org/10.2139/ssrn.4976449

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2025). AI companions reduce loneliness. *Journal of Consumer Research*. https://dx.doi.org/10.1093/jcr/ucaf040

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383. https://doi.org/10.2307/2666999

Epley, N., Kardas, M., Zhao, X., Atir, S., & Schroeder, J. (2022). Undersociality: Miscalibrated social cognition can inhibit social connection. *Trends in Cognitive Sciences*, *26*(5), 406–418. https://doi.org/10.1016/j.tics.2022.02.007

Epley, N., & Schroeder, J. (2014). Mistakenly seeking solitude. *Journal of Experimental Psychology: General*, *143*(5), 1980–1999. https://doi.org/10.1037/a0037323

Farrell, H., Gopnik, A., Shalizi, C., & Evans, J. (2025). Large AI models are cultural and social technologies. *Science*, *387*(6739), 1153–1156. https://doi.org/10.1126/science.adt9819

Folk, D., & Dunn, E. (2023). A systematic review of the strength of evidence for the most commonly recommended happiness strategies in mainstream media. *Nature Human Behaviour*, *7*(10), 1697–1707. https://doi.org/10.1038/s41562-023-01651-4

Glickman, M., & Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, *9*(2), 345–359. https://doi.org/10.1038/s41562-024-02077-2

Gonzalez, C., & Heidari, H. (2025). A cognitive approach to human–AI complementarity in dynamic decision-making. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-025-00499-x

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Gremler, D. D., & Gwinner, K. P. (2000). Customer-Employee Rapport in Service Relationships. *Journal of Service Research*, *3*(1), 82–104. https://doi.org/10.1177/109467050031006

Guilbeault, D., Baronchelli, A., & Centola, D. (2021). Experimental evidence for scale-induced category convergence across populations. *Nature Communications*, *12*(1), 327. https://doi.org/10.1038/s41467-020-20037-y

Gunaydin, G., Oztekin, H., Karabulut, D. H., & Salman-Engin, S. (2021). Minimal Social Interactions with Strangers Predict Greater Subjective Well-Being. *Journal of Happiness Studies*, *22*(4), 1839–1853. https://doi.org/10.1007/s10902-020-00298-6

Hamedani, M. G., Markus, H. R., Hetey, R. C., & Eberhardt, J. L. (2024). We built this culture (so we can change it): Seven principles for intentional culture change. *American Psychologist*, *79*(3), 384–402. https://doi.org/10.1037/amp0001209

Hecht, C. A., Ong, D. C., Clapper, M., Jones, M., Demszky, D., Yang, D., Eichstaedt, J. C., Bryan, C. J., & Yeager, D. S. (2025). Using large language models in behavioral science interventions: Promise & risk. *Behavioral Science & Policy*, *11*(1), 1–9. https://doi.org/10.1177/23794607251344698

Jackson, J. C., Castelo, N., & Gray, K. (2020). Could a rising robot workforce make humans less prejudiced? *American Psychologist*, *75*(7), 969–982. https://doi.org/10.1037/amp0000582

Jackson, J. C., Yam, K. C., Tang, P. M., Sibley, C. G., & Waytz, A. (2023). Exposure to automation explains religious declines. *Proceedings of the National Academy of Sciences*, *120*(34), e2304748120. https://doi.org/10.1073/pnas.2304748120

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. *Journal of Business Research*, *58*(4), 500–507. https://doi.org/10.1016/S0148-2963(03)00140-1

Kiesler, D. J. (1996). *Contemporary interpersonal theory and research: Personality, psychopathology, and psychotherapy* (pp. xviii, 398). John Wiley & Sons.

Konya, A., Thorburn, L., Almasri, W., Leshem, O. A., Procaccia, A., Schirch, L., & Bakker, M. (2025). Using collective dialogues and AI to find common ground between Israeli and Palestinian peacebuilders. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 312–333. https://doi.org/10.1145/3715275.3732022

Kozlowski, A. C., & Evans, J. (2025). Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. *Sociological Methods & Research*, *54*(3), 1017–1073. https://doi.org/10.1177/00491241251337316

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lai, S., Potter, Y., Kim, J., Zhuang, R., Song, D., & Evans, J. (2024). *Evolving AI collectives to enhance human diversity and enable self-regulation* (No. arXiv:2402.12590). arXiv. http://arxiv.org/abs/2402.12590

Macintosh, G. (2009). The role of rapport in professional services: Antecedents and outcomes. *Journal of Services Marketing*, *23*(2), 70–78. https://doi.org/10.1108/08876040910946332

Newman, M., Barabási, A.-L., & Watts, D. J. (2011). *The Structure and Dynamics of Networks*. Princeton University Press. https://doi.org/10.1515/9781400841356

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Oktar, K., & Lombrozo, T. (2025). How beliefs persist amid controversy: The paths to persistence model. *Psychological Review*. https://doi.org/10.1037/rev0000583

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* (No. arXiv:2304.03442). arXiv. https://doi.org/10.48550/arXiv.2304.03442

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). *Generative Agent Simulations of 1,000 People* (No. arXiv:2411.10109). arXiv. https://doi.org/10.48550/arXiv.2411.10109

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Reis, H. T., & Shaver, P. (1988). Intimacy as an interpersonal process. In *Handbook of personal relationships: Theory, research and interventions* (pp. 367–389). John Wiley & Sons.

Rubin, M., Li, J. Z., Zimmerman, F., Ong, D. C., Goldenberg, A., & Perry, A. (2025). Comparing the value of perceived human versus AI-generated empathy. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02247-w

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sandstrom, G. M., & Boothby, E. J. (2021). Why do people avoid talking to strangers? A mini meta-analysis of predicted fears and actual experiences talking to a stranger. *Self and Identity*, *20*(1), 47–71. https://doi.org/10.1080/15298868.2020.1816568

Sandstrom, G. M., Boothby, E. J., & Cooney, G. (2022). Talking to strangers: A week-long intervention reduces psychological barriers to social connection. *Journal of Experimental Social Psychology*, *102*, 104356. https://doi.org/10.1016/j.jesp.2022.104356

Sandstrom, G. M., & Dunn, E. W. (2014). Social interactions and well-being: The surprising power of weak ties. *Personality and Social Psychology Bulletin*, *40*(7), 910–922. https://doi.org/10.1177/0146167214529799

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, *5*(1), 46–57. https://doi.org/10.1038/s42256-022-00593-2

Shiiku, S., Marjieh, R., Anglada-Tort, M., & Jacoby, N. (2025). *The dynamics of collective creativity in human-AI hybrid societies* (No. arXiv:2502.17962). arXiv. https://doi.org/10.48550/arXiv.2502.17962

Smaldino, P. E., Moser, C., Pérez Velilla, A., & Werling, M. (2024). Maintaining transient diversity is a general principle for improving collective problem solving. *Perspectives on Psychological Science*, *19*(2), 454–464. https://doi.org/10.1177/17456916231180100

Smith, M. G., Bradbury, T. N., & Karney, B. R. (2025). Can generative AI chatbots emulate human connection? A relationship science perspective. *Perspectives on Psychological Science*, 17456916251351306. https://doi.org/10.1177/17456916251351306

Stewart, K. J. (2003). Trust Transfer on the World Wide Web. *Organization Science*, *14*(1), 5–17. https://doi.org/10.1287/orsc.14.1.5.12810

Sucholutsky, I., Collins, K. M., Jacoby, N., Thompson, B. D., & Hawkins, R. D. (2025). Using LLMs to advance the cognitive science of collectives. *Nature Computational Science*, *5*(9), 704–707. https://doi.org/10.1038/s43588-025-00848-z

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719), eadq2852. https://doi.org/10.1126/science.adq2852

Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025). Database Report: Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions. *Marketing Science*. https://doi.org/10.1287/mksc.2025.0262

Tracey, T. J. G., Ryan, J. M., & Jaschik-Herman, B. (2001). Complementarity of interpersonal circumplex traits. *Personality and Social Psychology Bulletin*, *27*(7), 786–797. https://doi.org/10.1177/0146167201277002

Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour*, *8*(10), 1864–1876. https://doi.org/10.1038/s41562-024-02001-8

Vesga, A., Van Leeuwen, N., & Lombrozo, T. (2025). Evidence for multiple kinds of belief in theory of mind. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001765

Williams, D. (2021). Socially adaptive belief. *Mind & Language*, *36*(3), 333–354. https://doi.org/10.1111/mila.12294

Wilson, D. S., Hayes, S. C., Biglan, A., & Embry, D. D. (2014). Evolving the future: Toward a science of intentional change. *Behavioral and Brain Sciences*, *37*(4), 395–416. https://doi.org/10.1017/S0140525X13001593

Yang, C. (Liu), Bauer, K., Li, X., & Hinz, O. (2025). My Advisor, Her AI, and Me: Evidence from a Field Experiment on Human–AI Collaboration and Investment Decisions. *Management Science*. https://doi.org/10.1287/mnsc.2022.03918

Yang, J., Zhang, M., Jin, Y., Chen, H., Wen, Q., Lin, L., He, Y., Xu, W., Evans, J., & Wang, J. (2025). *Topological structure learning should be a research priority for LLM-based multi-agent systems* (No. arXiv:2505.22467). arXiv. https://doi.org/10.48550/arXiv.2505.22467

Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). *The rise of AI companions: How human-chatbot relationships influence well-being* (No. arXiv:2506.12605). arXiv. https://doi.org/10.48550/arXiv.2506.12605