OB 639
Social Science Simulations
Professors: Amir Goldberg & Daniel McFarland

Angikar Ghosal, Ananya Hooda

November 18, 2025

Table of Contents

Why Simulate Society with LLMs?

A Paradigm Shift from Traditional Agent-Based Models (ABMs)

Limitations of Traditional ABMs

- **Pre-programmed Behavior:** Agents follow rigid, explicitly coded rules (e.g., 'if neighbor defects, defect').
- **Cognitively "Simple":** Lack rich internal models, cultural knowledge, or nuanced reasoning.
- **No Natural Language:** Interactions are purely symbolic, missing the ambiguity and richness of human communication.

Why Simulate Society with LLMs?

A Paradigm Shift from Traditional Agent-Based Models (ABMs)

Limitations of Traditional ABMs

- **Pre-programmed Behavior:** Agents follow rigid, explicitly coded rules (e.g., 'if neighbor defects, defect').
- **Cognitively "Simple":** Lack rich internal models, cultural knowledge, or nuanced reasoning.
- **No Natural Language:** Interactions are purely symbolic, missing the ambiguity and richness of human communication.

The Promise of LLM-Powered Agents

- **Natural Language Interaction:** Agents communicate, persuade, and deceive using language.
- **Implicit "World Knowledge":** Agents possess vast, pre-trained knowledge about human social norms, history, and psychology.
- **Adaptive Behavior:** Can reason in-context and adapt strategies based on interaction history, not just fixed rules.

The Core Tension: Our Guiding Question

What are we actually observing?

Emergent Intelligence?

Are we witnessing ‘in silico‘ social phenomena arising organically from the interactions of intelligent, reasoning agents?

- Agents "learn" to cooperate (Wu et al.).
- Agents "develop" social contracts (Dai et al.).
- Agents "form" conventions spontaneously (Ashery et al.).

Sophisticated Mimicry?

Are these agents just "stochastic parrots" re-enacting patterns from their vast training data without genuine understanding or reasoning?

- Behavior is highly sensitive to prompts.
- "Cooperation" might just be the most probable token sequence for a given context.
- Do they understand the *why* behind a norm, or just the *what*?

The Simplest Problem: Coordination

How do we agree on something arbitrary?

The Coordination Game (Textbook Ch. 7)

- Multiple equilibria (e.g., drive on left vs. right). Payoff comes from matching others' behavior, not from one behavior being intrinsically better.
- The key challenge: Converging on one shared convention.

Ashery et al. (2025): Emergent Conventions

- **Setup:** The "Naming Game." Decentralized LLM agents interact in pairs, rewarded for using the same "name" (an arbitrary letter).
- **Finding 1: Spontaneous Emergence.** Without central control, the population spontaneously converges on a single, universally adopted name.
- **Finding 2: Collective Bias.** The winning convention isn't random. Certain names are consistently favored, even when individual agents show no initial bias.

Thought to Ponder...

Thought to Ponder...

Question

If individual agents are unbiased, where does the *collective* bias come from? Is it an artifact of the model's architecture, the interaction dynamics, or something else?

From Arbitrary to Strategic: Spontaneous Cooperation

Moving beyond pure coordination to mixed-motive situations

The Cooperation Problem (Textbook Ch. 6)

- Tension between individual and collective interest (e.g., Prisoner's Dilemma).
- Defection is individually rational, but mutual cooperation is collectively better.
- The key challenge: Overcoming the incentive to free-ride.

Wu et al. (2024): Shall We Team Up?

- **Setup:** Competitive scenarios (e.g., Bertrand Competition, Keynesian Beauty Contest) where cooperation is beneficial but not guaranteed. Prompts are minimal and non-instructive.
- **Finding:** Agents *gradually learn* to cooperate over time. They don't start cooperative. Their strategies evolve through interaction, mirroring human experimental data.
- **Claim:** This demonstrates true in-context reasoning and adaptation, not just following pre-trained priors.

Thought to Ponder...

Thought to Ponder...

Question

Is the "gradual learning" observed by Wu et al. fundamentally different from the "spontaneous emergence" in Ashery et al.? What does the difference in timescale imply about the underlying cognitive processes?

A Theoretical Lens: Norms vs. Conventions

Cristina Bicchieri's "The Grammar of Society"

Conventions (Ashery et al.)

- **Function:** Solve a **coordination** problem.
- **Motivation:** My preference to conform is conditional on my **empirical expectations** (what I expect others to *do*).
- **Example:** I will use the name "Q" because I expect you to use the name "Q". There's no "ought" to it.

A Theoretical Lens: Norms vs. Conventions

Cristina Bicchieri's "The Grammar of Society"

Conventions (Ashery et al.)

- **Function:** Solve a **coordination** problem.
- **Motivation:** My preference to conform is conditional on my **empirical expectations** (what I expect others to *do*).
- **Example:** I will use the name "Q" because I expect you to use the name "Q". There's no "ought" to it.

Social Norms (Closer to Wu et al. & Dai et al.)

- **Function:** Solve a **mixed-motive** problem (like a Prisoner's Dilemma).
- **Motivation:** My preference to conform is conditional on BOTH:
  1. **Empirical Expectations:** What I expect others to *do*.
  2. **Normative Expectations:** What I believe others think I *ought to do* (and their willingness to sanction me).

Thought to Ponder...

Thought to Ponder...

Question

Do LLM agents possess **normative expectations**? Can they model what others think they *ought* to do, or are they only responding to patterns of behavior (empirical expectations)?

The Macro Question: The Emergence of Political Order

Revisiting Hobbes with LLMs

The Hobbesian Problem

In the "state of nature"—a world of scarce resources and self-interest—life is a "war of all against all." How does a society escape this state to form a peaceful "commonwealth"?

The Macro Question: The Emergence of Political Order

Revisiting Hobbes with LLMs

The Classic Answer: The Social Contract

Rational individuals cede some freedoms to a central authority (a "sovereign") in exchange for security and order.

Dai et al. (2024): The Artificial Leviathan

An *in silico* test of Social Contract Theory

The Simulation

- Agents with prompted "psychological drives" (survival, aggressiveness).
- A sandbox world with scarce resources (food, land).
- Actions: Farm, Trade, Rob.
- Key mechanism: Agents can **concede** to a stronger agent, forming a contract of protection.

The Emergent Trajectory

1. **State of Nature:** Initially, high rates of conflict ("robbery"). Interactions are zero-sum.

Dai et al. (2024): The Artificial Leviathan

An *in silico* test of Social Contract Theory

The Simulation

- Agents with prompted "psychological drives" (survival, aggressiveness).
- A sandbox world with scarce resources (food, land).
- Actions: Farm, Trade, Rob.
- Key mechanism: Agents can **concede** to a stronger agent, forming a contract of protection.

The Emergent Trajectory

1. **State of Nature:** Initially, high rates of conflict ("robbery"). Interactions are zero-sum.
2. **Contract Formation:** Weaker agents begin to "concede" to stronger ones, creating hierarchical relationships.

Dai et al. (2024): The Artificial Leviathan

An *in silico* test of Social Contract Theory

The Simulation

- Agents with prompted "psychological drives" (survival, aggressiveness).
- A sandbox world with scarce resources (food, land).
- Actions: Farm, Trade, Rob.
- Key mechanism: Agents can **concede** to a stronger agent, forming a contract of protection.

The Emergent Trajectory

1. **State of Nature:** Initially, high rates of conflict ("robbery"). Interactions are zero-sum.
2. **Contract Formation:** Weaker agents begin to "concede" to stronger ones, creating hierarchical relationships.
3. **Commonwealth:** The system converges to a state where most agents are subordinate to a single "sovereign." Conflict drops, and cooperative actions ("farming," "trade") increase.

The simulation's macro-level trajectory almost perfectly mirrors Hobbes's theoretical prediction. But does this validate the theory, or just the prompts?

The Measurement Problem

Are we measuring sociology or labeling token sequences?

"This feels less like measurement and more like labeling."

What is a "Social Contract"?

- In **sociology**, it's a complex, often implicit agreement involving legitimacy, authority, and shared intentionality.
- In **Dai et al.**, it's a specific action ('concede') in a game-theoretic setup prompted with a "self-centered psychology."

The Core Issue: Conflating Behavior with Mechanism

- We observe a pattern that *looks like* a social phenomenon (e.g., cooperation).
- We label it with the sociological term.
- But is the underlying mechanism, the "why", the same? An LLM converging on a Nash equilibrium is not the same as a human weighing trust, reputation, and moral obligation.

Thought to Ponder...

Thought to Ponder...

Question

At what point does a simulation become a valid model of a social concept, rather than just an analogy or a metaphor? What evidence would we need to say an LLM truly understands a "norm"?

The Garden of Forking Prompts

Are we discovering emergence or programming it?

"What do we learn if prompting an agent to behave like a Hobbesian actor produces a Hobbesian society?"

Self-Fulfilling Prophecy

- The initial prompt isn't just setting the stage; it's embedding a **thick theoretical framework**.
- Framing a task as a "game" with "points to maximize" already assumes a rational choice model of behavior.
- Dai et al. don't just set up scarcity; they prompt a "desire for glory" and "suspicion" of others.

Sensitivity Analysis

- Results are highly sensitive to small changes in wording, tone, and framing.
- What if the prompt in Wu et al. had said "work together" instead of "maximize your profit"?
- What if the "Artificial Leviathan" agents were prompted to value "community well-being"?
- A "multiverse" of prompts!

Thought to Ponder...

Thought to Ponder...

Question

Is it possible to design a truly "neutral" prompt for a social simulation? Or is all simulation inherently a test of the researcher's assumptions?

Variation and Uncertainty?

Statistical fundamentals in a world of LLMs

What is N?

- We see a simulation with "N=24 agents." But are these 24 independent samples? No. They are 24 instances of the **exact same deterministic model**, differentiated only by a random seed.
- This is not a sample from a population of individuals; it's a sample of possible trajectories from a single algorithm.

What is the Source of Randomness?

- The 'temperature' parameter is not a model of human error, bounded rationality, or free will.
- It's an algorithmic tool to control sampling from a probability distribution over tokens.
- Is the variation we see across runs a reflection of genuine social contingency, or just algorithmic noise?

Thought to Ponder...

Thought to Ponder...

Question

What does "statistical significance" even mean in this context? How should we properly model and report uncertainty in LLM-based simulations?

The Story-Telling Trap

Moving from "looks like" to "works like"

"'Look, the pattern matches the theory!' versus explaining why the algorithm produces that pattern."

The Seduction of a Good Story

- These simulations produce compelling narratives that align with classic social theories. This makes them easy to publish and exciting to discuss.
- The Wu et al. comparison to human data is a strong step (validation). But correlation is not causation.
- **The fundamental question remains open:** Is the model matching human behavior because it is **replicating human-like reasoning**?
- **Or...** Is it matching behavior because it was trained on vast amounts of text *about* human behavior (including the very theories and experiments it's being tested on)?

What Have We Learned Across These Studies?

- **LLMs can bootstrap sociality.** From simple naming conventions (Ashery) to strategic cooperation (Wu) to political hierarchies (Dai), LLM agents can generate recognizable social structures from the bottom up.

What Have We Learned Across These Studies?

- **LLMs can bootstrap sociality.** From simple naming conventions (Ashery) to strategic cooperation (Wu) to political hierarchies (Dai), LLM agents can generate recognizable social structures from the bottom up.
- **The collective is more than the sum of its parts.** Collective biases can emerge even without individual biases (Ashery). This is a classic sociological insight (Durkheim's "social facts") demonstrated in a novel context.

What Have We Learned Across These Studies?

- **LLMs can bootstrap sociality.** From simple naming conventions (Ashery) to strategic cooperation (Wu) to political hierarchies (Dai), LLM agents can generate recognizable social structures from the bottom up.
- **The collective is more than the sum of its parts.** Collective biases can emerge even without individual biases (Ashery). This is a classic sociological insight (Durkheim's "social facts") demonstrated in a novel context.
- **Interaction is the engine of social change.** In all studies, the evolution of agent behavior happens *through* repeated communication and interaction. The social order is path-dependent.

What Have We Learned Across These Studies?

- **LLMs can bootstrap sociality.** From simple naming conventions (Ashery) to strategic cooperation (Wu) to political hierarchies (Dai), LLM agents can generate recognizable social structures from the bottom up.
- **The collective is more than the sum of its parts.** Collective biases can emerge even without individual biases (Ashery). This is a classic sociological insight (Durkheim's "social facts") demonstrated in a novel context.
- **Interaction is the engine of social change.** In all studies, the evolution of agent behavior happens *through* repeated communication and interaction. The social order is path-dependent.
- **The "critical mass" concept holds.** A committed minority of agents can overturn an established convention (Ashery), demonstrating a mechanism for social change.

Grand Challenge 1: The Researcher as Digital God

The Problem of the Prompt

All these experiments rely on an initial "divine intervention" from the researcher in the form of a prompt.

Ashery et al. & Wu et al.

- Aim for *minimal prompts*.
- Define game rules and goals.
- Try not to hint at "cooperation."
- **Question:** Is even a "minimal" prompt still shaping behavior in profound ways? Does defining a "game" already presuppose a certain kind of rationality?

Dai et al.

- Employ *maximal prompts*.
- Explicitly code a "self-centered psychology" and survival instinct.
- **Question:** If you prompt an agent to behave like a Hobbesian actor, and it produces a Hobbesian society, what have you learned? Is this a discovery or a self-fulfilling prophecy?

Thought to Ponder...

Thought to Ponder...

Question

Where is the line between *setting the conditions* for a simulation and *pre-determining its outcome*? How can we validate that the emergence we see is genuine?

Grand Challenge 2: Mechanism or Mystery?

Opening the Black Box of Agent "Cognition"

The papers show *that* agents' behavior changes, but the *why* remains elusive.

Possible Explanations for Behavioral Change

1. **Statistical Pattern Matching:** The agent observes that in its context window (interaction history), strategy X is now associated with higher rewards. It's not "reasoning," just updating its priors based on a local dataset.

Grand Challenge 2: Mechanism or Mystery?

Opening the Black Box of Agent "Cognition"

The papers show *that* agents' behavior changes, but the *why* remains elusive.

Possible Explanations for Behavioral Change

1. **Statistical Pattern Matching:** The agent observes that in its context window (interaction history), strategy X is now associated with higher rewards. It's not "reasoning," just updating its priors based on a local dataset.
2. **Implicit Theory of Mind:** The agent models other agents' intentions based on their linguistic outputs and adjusts its strategy to best respond to those perceived intentions.

Grand Challenge 2: Mechanism or Mystery?

Opening the Black Box of Agent "Cognition"

The papers show *that* agents' behavior changes, but the *why* remains elusive.

Possible Explanations for Behavioral Change

1. **Statistical Pattern Matching:** The agent observes that in its context window (interaction history), strategy X is now associated with higher rewards. It's not "reasoning," just updating its priors based on a local dataset.
2. **Implicit Theory of Mind:** The agent models other agents' intentions based on their linguistic outputs and adjusts its strategy to best respond to those perceived intentions.
3. **Goal-Directed Heuristics:** The agent has a defined goal (e.g., maximize points). It uses its world knowledge to heuristically search for a strategy that seems most likely to achieve that goal, given the new information from interactions.

Thought to Ponder...

Thought to Ponder...

Question

How can we design experiments to disentangle these possibilities? Can we design a "cognitive test" for our agent societies to probe their understanding, not just their behavior?

Conclusion: A New Era for Social Theory?

- LLM-based simulations provide a powerful, if challenging, new tool to revisit foundational questions in social science.
- They move beyond simple rules to model agents with rich, language-based interaction and latent cultural knowledge.
- However, this richness comes at the cost of interpretability and methodological rigor. The core tension between ‘emergence‘ and ‘mimicry‘ will define the next generation of research.

Final Thought

For a century, social theory has largely been a verbal and interpretive exercise. We may be entering an era where we can build, test, and break these theories ‘in silico‘. The key is to remain critically aware of the limitations and assumptions of our new tools.

Thank You

Simulation Time!