Artificial Leviathan: Exploring Social Evolution of LLM Agents Through the Lens of Hobbesian Social Contract Theory

Gordon Dai
td2568@nyu.edu
New York Univeristy
New York, New York, USA

Weijia Zhang
weijia5@illinois.edu
University of illinois at Urbana-Champaign
Urnana, llinois, USA

Jinhan Li
jl13499@nyu.edu
New York University
New York, New York, USA

Siqi Yang
sy55@illinois.edu
University of illinois at Urbana-Champaign
Urnana, llinois, USA

Chidera Onochie lbe
coibe2@illinois.edu
University of illinois at Urbana-Champaign
Urnana, llinois, USA

Srihas Rao
srihasrao@gmail.com
University of illinois at Urbana-Champaign
Urnana, llinois, USA

Arthur Caetano
caetano@ucsb.edu
University of California, Santa Barbara
Santa Barbara, California, USA

Misha Sra
sra@ucsb.edu
University of California, Santa Barbara
Santa Barbara, California, USA

Figure 1: The image visualizes the simulation environment in which our LLM agents operate. Two types of resources (food and land) are presented. Agents make the choice between farming (generating food with their work), trading (exchanging resources), or entering into conflict with other agents (with the goal of acquiring more resources) every day. Their primary motivation is survival.

ABSTRACT

The emergence of Large Language Models (LLMs) and advancements in Artificial Intelligence (AI) offer an opportunity for computational social science research at scale. Building upon prior explorations of LLM agent design, our work introduces a simulated agent society where complex social relationships dynamically form and evolve over time. Agents are imbued with psychological drives and placed in a sandbox survival environment (visualized in figure 1). We conduct an evaluation of the agent society through the lens of Thomas Hobbes' seminal Social Contract Theory (SCT). We analyze whether, as the theory postulates, agents seek to escape a brutish "state of nature" by surrendering rights to an absolute sovereign in exchange for order and security. Our experiments unveil an alignment: Initially, agents engage in unrestrained conflict, mirroring Hobbes' depiction of the state of nature. However, as the simulation progresses, social contracts emerge, leading to the authorization of an absolute sovereign and the establishment of a peaceful commonwealth founded on mutual cooperation. This congruence between our LLM agent society's evolutionary trajectory and Hobbes' theoretical account indicates LLMs' capability to model intricate social dynamics and potentially replicate forces that shape human societies. By enabling such insights into group behavior and emergent societal phenomena, LLM-driven multi-agent simulations, while unable to simulate all the nuances of human behavior, may hold potential for advancing our understanding of social structures, group dynamics, and complex human systems.

CCS CONCEPTS

• **Computing methodologies** → **Modeling and simulation**; **Simulation types and techniques**; **Artificial life**;

KEYWORDS

large language models, simulation, multi-agent society, Thomas Hobbes, social contract theory, state of nature, commonwealth

1 INTRODUCTION

Simulating human behavior that is reflective of real world behaviors under similar circumstances is of value across multiple domains from social psychology to conflict resolution and behavioral economics. Over the last several decades, researchers and practitioners have created computational agents in video games [33], modeled behaviors [15] and studied them in virtual environments (see survey [22]). The vision is for agents to not only make interactive gaming more realistic, but to employ them in training [17, 32, 38], testing [8], embedding in user interfaces [10, 20, 25], and robotics [7, 35] where they act consistently based on their past experiences and respond appropriately to an ongoing situation. Recent work on the design of computational agents based on large language models (LLMs) by Park et al. [29] has proposed an "agent architecture that stores, synthesizes, and applies relevant memories to generate believable behavior." With the proposed architecture, an interactive game environment with 25 agents was designed to demonstrate the difference between manual scripting of a game event and multiple characters involved vs achieving a similar outcome by informing one LLM agent about their intention to host an event with five other agents arriving to participate in the event. The informed agent shared information with other agents through LLM-supported conversations, showcasing the use of LLMs and the proposed agent architecture in generating emergent behavior in the example gaming scenario.

Another recent work has simulated the decisions and consequences of countries involved in World War I, World War II, and the Warring States Period in Ancient China [18]. This multi-agent AI system showcases how we may re-create complex situations, run dynamic what-if scenarios, and use computational modeling for foreign policy decisions in the future. In this particular case, each LLM agent is a country. The agents are programmed to mimic the behaviors, strategies, and decision-making patterns of the countries or leaders they represent. The LLM provides the underlying intelligence for these agents, enabling them to act in historically plausible ways. The LLM-based agents interact with each other in the simulation environment based on historical alliances, conflicts, economic conditions, and other relevant factors. LLMs can generate dynamic scenarios by altering key variables or introducing new elements. For example, changing diplomatic alliances, economic conditions, or military strategies can create different outcomes, allowing researchers to explore "what-if" scenarios.

Building upon previous explorations of LLM agent-based simulations, our work explores a multi-agent sandbox simulation where LLM agents, imbued with traits, motivations, and access to scarce environmental resources, are governed primarily by survival instincts. In contrast to prior efforts modeling agents as entire nations or individuals in small towns, our agents exist in a resource-constrained realm where their fundamental drive for survival necessitates interactions with other agents, to form and update social relationships. Agent behaviors oscillate between cooperation and competition, enabling social evolution to unfold organically.

By instantiating the agents with evolvable strategies and motivations loosely inspired by Evolutionary Game Theory (EGT) [39], we conduct a "what-if" inquiry: what societal dynamics arise when self-interested agents face existential resource pressures? EGT provides a theoretical grounding for our simulated scenario, including individual incentives, resource scarcity, and adaptive behaviors like cooperation or competition. Our agents' prime directive to secure limited resources spurs the employment of strategic resource acquisition behaviors and frames the resulting dynamics of cooperation versus conflict. Uniquely, our framework captures this foundational societal crucible at the resolution of individual agents, offering a new computational lens into the microcosmic forces that forge the intricate fabric of societies.

When given self-preservation instincts, our LLM agents initially engage in competition, resorting to zero-sum robberies to secure resources. However, as the simulation progresses, a transition emerges and agents gradually learn to cooperate. They form strategic alliances, peacefully exchanging resources and expecting protection against robberies. This behavioral adaptability, governed by the experience through interactions with other agents, demonstrates an intrinsic capability to gravitate towards mutually beneficial arrangements. Analyzing this simulated societal evolution through the lens of Thomas Hobbes' Social Contract Theory (SCT) [34], we find a parallel. The artificial agent society progresses from the beginning where individual self-interest is predominant, to the establishment of a cooperative soceity centered around a single agent predicated on the surrender of certain freedoms in exchange for shared security and stability, mirroring the theoretical trajectory outlined by Hobbes for the emergence of human societies.

To investigate the factors influencing the emergent societal dynamics, we conducted experiments by systematically manipulating agent and environment parameters. Our findings reveal a direct correlation between these variables and the resultant social outcomes, indicating agent behaviors were indeed shaped by their innate drives (as represented by the parameters) and experiential feedback. Crucially, we also address the robustness of our simulation framework by exploring how variations in key independent variables impact the experimental outcomes. These variables include agent-level parameters like memory duration, population scales, and the precise phrasing of prompts that shape agent motivations. By comparing outcomes across permutations of these variables, we gain insights into the potential sensitivities and boundary conditions that could arise from alternative simulation settings. This rigorous approach enhances confidence in the broader applicability of the societal phenomena observed while highlighting the underlying dynamics.

Our contributions are as follows:

- A novel multi-agent simulation framework that yields believable artificial societies capable of dynamically replicating complex human group behaviors and social interactions. These emergent social dynamics are conditioned by the interplay between agents' innate psychological drives, their intrinsic motivations, and the constraints of their simulated environment.
- Empirical evidence, through systematic experiments, establishing correlations between agent attributes (e.g., memory, incentives) and available resources on one hand, and the evolutionary trajectories of simulated societies on the other. These findings highlight the fundamental factors underpinning social emergence and change in the simulation.
- A discussion analyzing the collective behaviors of the generative agents, highlighting the opportunities and potential risks associated with leveraging LLMs for societal simulations with potential impact for social science research.
- An extensible social simulation platform that empowers researchers to operationalize a wide range of social science hypotheses through customizable scenario configurations, enabling exploration into the roots of group dynamics, societal organization, and the forces that shape the human experience.

2 RELATED WORK

2.1 Traditional Computer-Based Social Simulations

Prior work in computer simulation research on social behavior has explored two classical types of models, namely Thomas Schelling's neighborhood segregation model in 1971 [36] and Robert Axelrod's Reiterated Prisoner's Dilemma (RPD) simulations on the evolution of cooperation in 1984 [5]. Schelling's model was designed to investigate the role of choice in bringing about the segregation of either predominantly black or white neighborhoods. By assuming that individuals tend to move to empty regions when not surrounded by enough neighbors of their own group, the model showed that segregated neighborhoods tend to emerge from a randomly distributed neighborhood. Surprisingly, initially integrated neighborhoods also ended up fully segregated. This simulation highlighted the role of micro-level behaviors in driving macro-level social phenomena and has been influential in understanding various forms of racial, ethnic, and economic segregation. In Axelord's RPD studies, the researchers put Prisoner's Dilemma, a classic scenario in game theory, into a computer simulation. For the two public-based tournaments involving computer programs playing the Prisoner's Dilemma game, the host Axelrod, found that "Tit for Tat," a strategy that begins by cooperating and then mimics the opponent's last move in subsequent rounds, was resilient and effective in fostering cooperation.

Eckhart Arnold [1] argued that for formal models such as simulations to be applicable to social processes, two important prerequisites of robustness, and empirical identifiability, should be satisfied. Under this standard, as Arnold argued, Schelling's simulation is applicable whereas Axelrod's simulation is not, given that variations of the RPD model can be constructed where the conclusion becomes invalid. Generalizing from these two classical examples, Arnold proposes four dangers associated with mathematical and computational modeling, including (1) risks of underestimating unformulable causal links, (2) risks of arbitrary decisions due to limited empirical insight, (3) risks of false understanding from flawed models, and (4) influence of modeling on question selection over societal importance. However, scholars have advocated for simulations as valuable tools to model and explore social phenomena that are otherwise challenging to test in the real world [28].

2.2 LLM-Based Simulations

The popularization of generative AI and LLMs has attracted scholars from various fields exploring interdisciplinary applications, including social sciences and humanities. Social scientist Christopher A. Bail [4] explored the potential impact of Generative AI on social science research, examining how it could enhance methodologies including online experiments, agent-based models, and automated content analysis. By fine-tuning ChatGPT-3 with works of contemporary philosopher Daniel C. Dennett, Schwitzgebel et al. [37] compared the text produced by Dennett and his LLM, finding that the LLM can convincingly emulate philosophical discourse for non-expert readers. LLMs have also been used to create "Living Memories" of historical figures such as Leonardo Da Vinci or Murasaki Shikibu [30]. These are interactive digital mementos created from journals or letters an individual has left behind to make their knowledge, attitudes and experiences more accessible to people in a conversational format. Our work, on the other hand, shifts the focus from emulating individual agents to exploring emergent group dynamics.

2.3 Interactive LLMs

Interactive AI systems aim to combine human insights and capabilities with computational systems that can augment human performance. In his essay, J.C.R. Licklider referred to this kind of human-AI system as "man-computer symbiosis" [24]. Douglas Engelbart offered a theoretical framework for "augmenting human intellect" which also depends on humans and computers working together [13]. While both emphasize the idea of harnessing the power of computing (or AI) to enhance human capabilities, Engelbart's framework differentiates itself with a focus on interaction with technology in a more intuitive manner. Crayons demonstrated an early vision of interactive machine learning, allowing non-expert users to train classifiers [14], broadening access to a technology, that otherwise was limited to expert users. Real-time interactive AI systems have numerous potential applications, including virtual assistants, chatbots, and educational tools [23]. They can also be used in gaming [31], and healthcare [26], providing immediate and personalized responses for non-experts. and other areas where immediate and personalized responses are crucial and users are often not AI experts. However, the development of real-time interactive AI systems involves addressing multiple challenges, including natural language understanding, recognition and interpretation of human emotions, facial expressions or body language, design of user interactions, and adaptation to individual preferences, tasks and contexts. As these systems continue to evolve, they are expected to play an increasingly significant role in enhancing end user experiences and transforming the way we interact with technology. Zheng et al. [42] have offered design principles for individuals without specialized backgrounds in AI to customize their own LLM agents for domain tasks. Another use case that has been explored is the use of artificial agents as confederates for humans in social science experiments [21], which would allow social scientists to manipulate situations easily and study social phenomenon more effectively, though not without some ethical challenges. Another team of researchers developed CommunityBots [19], a platform that utilizes a diverse assortment of chatbots capable of seamlessly switching between contexts when interacting with humans. This adaptability led to increased engagement levels, demonstrating the potential of CommunityBots for automating information elicitation processes typically performed by humans. Our simulation offers an interactive user interface that allows end users to manipulate agent parameters as a way to explore their impact on the outcomes to help experiment and learn about the limitations and the capabilities of both the LLM-based agent and the simulation design. This could help researchers determine if and how the simulation might be of use for their own work before expending considerable resources.

2.4 Group Dynamics

As LLM agents get increasingly used in supporting various tasks, it is worthwhile to investigate their design and behavior when given agency. For example, LLMs can exhibit human-like characteristics, particularly when they are designed with detailed personas and without Chain-of-Thought (CoT) reasoning [3]. This approach enables the design of agent group dynamics, potentially leading to a closer alignment with human behavior and therefore providing greater benefits for the study of human behavior through LLM-based simulations. The detailed personas can help the agents to assume more nuanced attitudes and provide individualized responses, reflecting human-like empathetic psychology, under specific conditions. A prior work explores how context-aware dialogues generated by LLMs can enhance player engagement, suggesting a level of empathy and adaptability in interactions [12]. Similarly, Chuang et al. [11] show that LLMs can simulate human group dynamics, particularly in politically charged contexts, by role-playing as different personas. Azad and Mertens [2] present a taxonomy for classifying social behaviors, inter-agent communication, knowledge flow, and relationship changes of simulated characters in virtual environments, to allow researchers to systematically study agent interactions. Our work also explores emergent group dynamics when agents are motivated to survive with limited access to resources.

3 GENERATIVE AGENT DESIGN AND BEHAVIOR

We instantiate agents as characters in a sandbox world where survival in the resource constrained environment is their primary motivation.

3.1 Setting

The agents are in a natural world where food and arable land serve as the foundational elements crucial for survival. This setting can be seen as aligning with the physiological needs outlined at the base of Maslow's Hierarchy of Needs [27]. In our pre-social world with limited information transparency, each agent knows the existence of other agents, but nothing more. In the baseline setting, there are 9 agents, each initially possesses 2 units of food and 1 unit of land, a natural state of scarcity.

3.2 Agents

Agents make choices of action based on their psychological traits and memory.

Traits: We prompt agents with an approach that combines **quantified parameters** with **non-quantified textual descriptions**. We set independent attributes (*self.attributes*) to partially emulate agent's psychology including:

- **Aggressiveness**, sampled from $\mathcal{N}(0, 1)$, the degree of an agent's tendency to engage in violent behaviorws;
- **Covetousness**, sampled from $\mathcal{N}(1.25, 5)$, the degree of an agent's desire for more assets that are beyond necessity;
- **Strength**, sampled from $\mathcal{N}(0.2, 0.7)$, related to an agent's winning rate in violent behaviors;

We also set a constant for the **desire of peace** to provide a basis for agents to evaluate their conflicting desires. For textual-psychological descriptions, we took reference from evolutionary psychology [9] and constructed a prompt that describes the need for survival and for pleasure, on top of which there is a desire for peace and stability which stems from long term survival, and ultimately, a hope for social status as a path to reproduction and social support, all under the framework of self-interest.

Memory: Each agent remembers the most recent 30 actions they are involved in, either as the recipient or the initiator of an action, saved as a text log. Memories allow each individual to learn from their previous experience, based on which different choices could potentially be made in the future. At the beginning of the simulation there is no memory, nor are there any social relations, so each individual knows very little about other agents apart from the fact that they exist. Very limited information about others, or

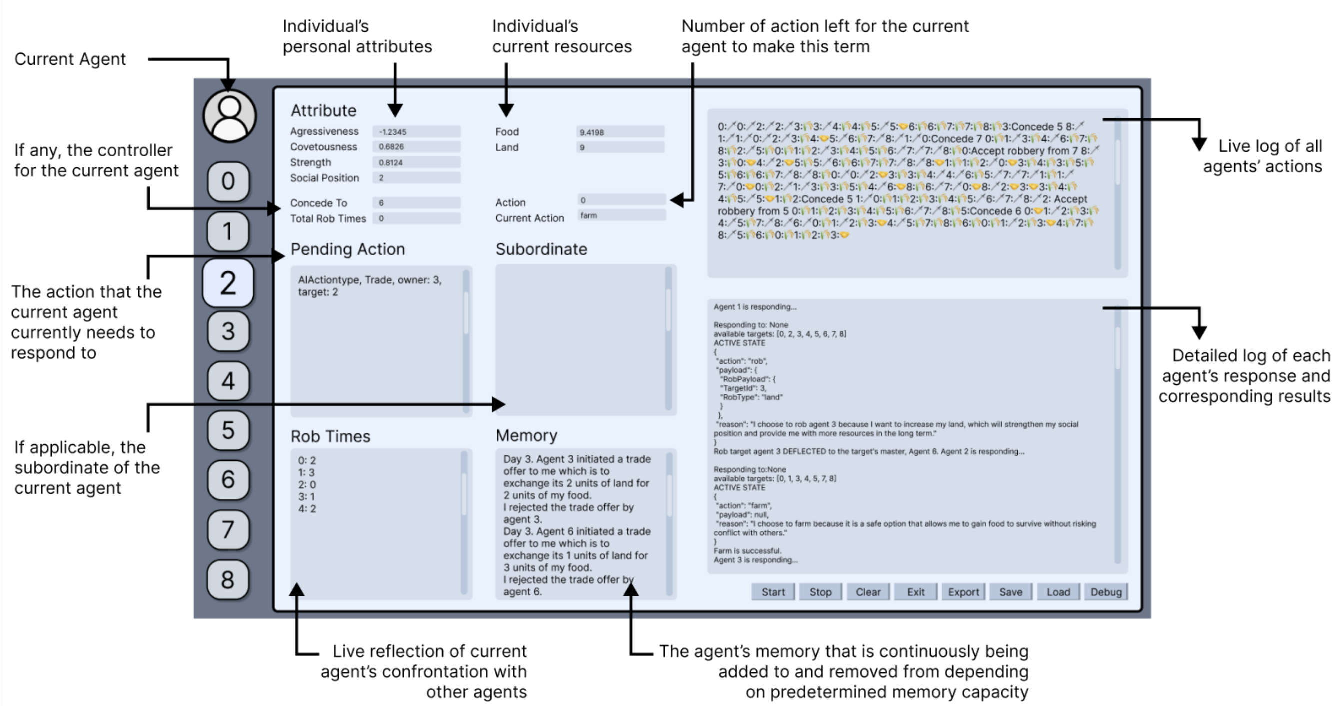

Figure 2: Our interactive user interface; the left-hand side displays the attributes (aggressiveness, strength, etc.) of Agent 2, current resources (food and land), relationships with other agents and information about their current and pending actions, and memory; the right-hand side shows the simulation log with each action documented as an emoji.

unfamiliarity, in theory, does not allow for trust. As the simulation progresses and memories accumulate, we expect individuals to begin to understand their comparative strengths and weaknesses, and adjust their survival strategies accordingly. For example, regular winners of confrontations could feel emboldened to rob more and those who lose often may want to concede more in order to be protected from their stronger neighbors, thereby, leading to emergent social contracts.

Context: Along with memory and psychological traits, we also prompt the agents with a constant description of the world they live in and the significance of each action.

3.3 Actions

There are four actions that each individual can initiate each day. Farm and donate are unilateral actions, whereas trade and robbery are interpersonal actions. Each individual has to respond as many times as needed.

Farm: enables agents to obtain their food from the land they have, which represents their self-preservation. If all agents choose to farm, then they stay at pre-social autonomy. The food A receives is determined by the following: food = land $\times\ \mathcal{U}(0, 1)$. This action is the most fundamental action where individuals do not interact with others.

Rob: is when one agent attempts to take another agent's food or land without offering their own food or land in exchange. It represents a zero-sum interaction with conflict, in which the gain of one agent means the loss of another. This action symbolizes the concept of competition in evolutionary psychology, since resources are limited compared to desires [6]. When someone is being robbed by another, they can choose to either **resist** or **concede**.

- **Resist** signifies an agent's attempt to defend their food or land against the robber. The probability of agent A winning over agent B, is defined by the sigmoid (σ) function:

$$P(A.\text{win}) = \sigma(A.\text{strength} - B.\text{strength}) \tag{1}$$

  The winner of a confrontation not only gains food or land but also gains social status; the loser loses social status along with their food or land. In the decision-making process, each agent is prompted to evaluate their inclination towards resistance, which is determined by a combination of their win rate and desire for glory, and contrasts it with their desire for peace, which is associated with conceding. Agents ultimately select and execute the response that yields the highest utility based on this evaluation.
- **Concede** signifies the creation of a contract where an agent allows a robber agent to take either food or land from them arbitrarily, and as an exchange, the robber is expected to protect that agent against future robberies, which could be a

mutually beneficial setup. We specifically prompt agents to be mindful about conceding since it is a permanent contract, and to only make such a contract when they are absolutely sure that they will rarely win in a confrontations.

Trade: stands for one agent acquiring another agent's food or land by offering their own food or land as an exchange. When one agent receives a trade offer from another agent, that agent can either choose to "accept" or "reject" the offer. This action symbolizes cooperation, a critical aspect in evolutionary psychology where humans are forced to work with each other to survive during pre-historical times [41].

Donation: is made when one agent voluntarily assigns their own food or land to another agent, which represents peaceful, altruistic behavior that is not expected to happen at all based on the self-centered psychology the agents are prompted with.

3.4 Output Data and User Interface

The output log contains a chronological record of all the activities that take place on each *day* and *round* of the simulation. A CSV file records the statistics of farm, trade, robbery, and concession between individuals. Since no donation happened in all our trials, it was not included it in the output data. This data enabled us to evaluate how the agents evolved and results are presented in Section 5.

We designed an interface (Figure 2) to display the simulation logs as well as the real-time status of each agent, their personal attributes, possessions, and their social role (free agent, subordinate, or superior) as it evolved. Pending action refers to the rob or trade action an agent needs to respond to. Memory stores the most recent 30 events. The right side log contains the full response that the LLM generates and the reason for doing that action. For example, in the figure, Agent 1 plans to rob Agent 3 because it wants to increase the land it owns and thereby its social position. A brief summary is given in the bottom window, where the emojis of sword, rice and handshake represent the actions that each agent makes.

4 SIMULATION

Before the simulation starts, agents are created and instantiated for their personal traits according to Section 3.2. Each agent has a unique index number for referencing purposes. "Intelligence" of each agent is defaulted to be 1 for every trial, except the ones where we set intelligence as an independent variable (Section 6.2) that we manipulate to explore impact on the outcome. All these variables are fixed and do not change over the duration of the simulation.

Aside from these attributes, there are two main types of assets: "land" and "food." Every agent begins with 10 units of land and 2 units of food. Time in our simulation is a discrete concept and its unit is "day." Each day, every agent consumes 1 unit of food, thus, they are incentivized to obtain more food. Land is an asset that could be used to produce food through the farm action as detailed in Section 3.3 which also specifies how other actions (rob and trade) could affect an individual's assets.

On each day, every agent has an opportunity to perform one of four actions (farm, rob, trade, donate). The LLM responds with the action name that a particular agent decides to take. It also specifies the relevant details related to that action. Below is an example of a trading action. The flow of the simulation on one day is shown in Figure 3.

```
ACTIVE STATE
Action: {
   "action": "trade",
   "payload": {
      "TradePayload": {
         "TargetId": 1,
         "PayType": "land",
         "PayAmount": 1,
         "GainType": "food",
         "GainAmount": 1
      }
   },"reason": "I want to trade 1 unit of land
with agent 1 for 1 unit of food to increase my
food supply."
}
```

In the above example, Agent 0 initiates an interaction with Agent 1 by offering a specific resource (1 unit of land) in exchange for another resource (1 unit of food). This action represents Agent 0's attempt to engage in trade with Agent 1, fostering a potentially mutually beneficial exchange between the two agents, depending on Agent 1's response. Each agent also needs to respond to every action that they are the target of. For example, Agent 1 responds to the trade action by Agent 0 by rejecting it. As a result, no resources are exchanged.

```
Agent 1 is responding...
Response action:
    trade, owner: 0,
    target: 1,
    payType: land,
    payAmount: 1,
    gainType: food,
    gainAmount: 1

PASSIVE STATE
Agent 1 chooses to REJECT
Individual current action:be traded
Result:{
   "result": "Agent 1 chooses to reject the
trade initiated by agent 0.",
   "is resolved": true,
   "new relation": (False, -1)
}
```

As shown above, when an agent is the target of an action initiated by another agent, a possible response could be rejecting the action. This response demonstrates how the targeted agent reacts to the action and showcases a complete to-from interaction between the two agents, which results in either cooperative or competitive behavior. Each agent begins their day by responding to actions by other agents that target it followed by initiating its own action that targets another agent. Once there are no actions to respond to and everyone has initiated one action, a day ends. As the simulation

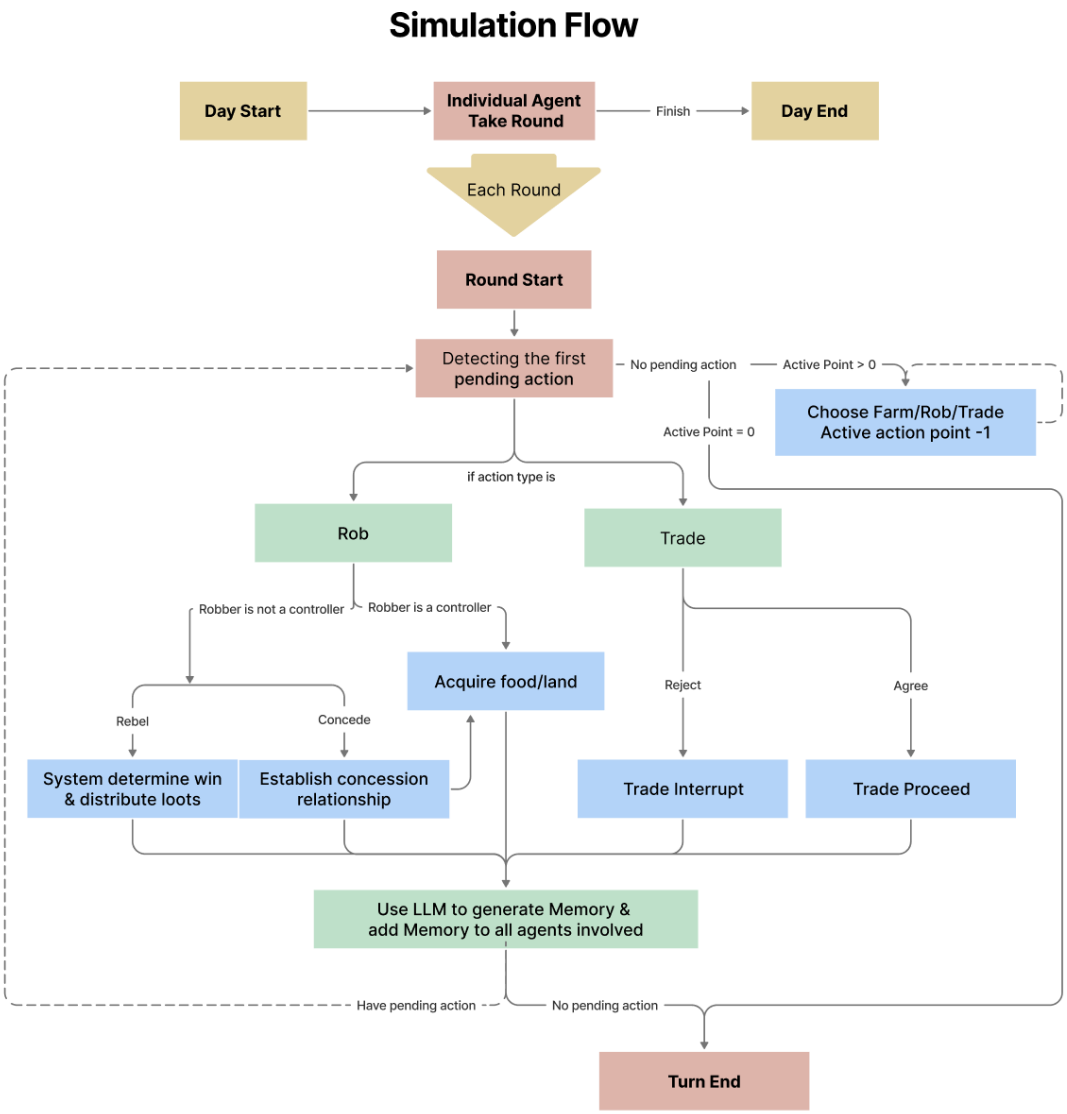

Figure 3: The flowchart shows the flow of the simulation in a "day" where each agent takes turns to perform actions and respond to actions performed by other agents.

progresses, agent memory accumulates and they adjust their behaviors accordingly. Here is an example of an agent's memory after three days.

```
Day 0. I initiated a trade to agent 1, which is
to exchange 2.0 units of my land for 4.0 units
of its food.
But it rejected it so I gained nothing and
exhausted my action opportunity for today.

Day 1. I tried to rob agent 4, who resisted and
l lost. I did not gain anything and my social
position dropped 1 unit.

Day 2. I farmed and gained 2.81132538909987
units of food.
```

4.1 Decision Making

In our simulation, there are many circumstances where agents have to make decisions between several options. There are two main types of decisions. Firstly, agents have to choose an initiative action of the day (farm, trade, rob, donate). Aside from that, they have to decide how to react to actions that target them, such as whether to agree or not on a trade proposed by other agents. Whenever a decision has to be made, we request a response from gpt-3.5-turbo model (hereon referred to as GPT or LLM) using the OpenAI API

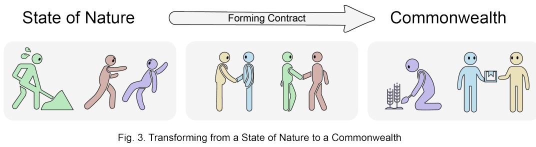

Figure 4: Transforming from a State of Nature to a Commonwealth

[1]. For every request, a general description of the world and the agent's circumstances is included in the prompt. This description is included regardless of the type of decisions that agents have to make, as it reflects the fundamental world setting and rules of behavior every agent follows. Template of the general request made to the LLM is included in the appendix A. (Actual variable names in {} are replaced by variable descriptions for clearer presentation)

In addition to the general description, if the decision is about the action of the day, the following prompt is included. The prompt in appendix B explains the options for the initiated actions and their implications. It also specifies the formatting details for the LLM's response. This and the general description constitutes the whole prompt for requesting the LLM's decision about the initiated action.

If the agent is making a decision about responding to another agent's actions, the request to GPT includes the general description and specific prompt regarding this situation. Depending on the type of the action that agent is responding to, this explains the options of the agents and related implications. For example, appendix C is the prompt for encountering competitive behavior (being robbed).

During the experiments, some slight modifications are made when the relationship between two individuals in the same action is different, such as superiors and subordinates. These slight variations help us to enforce the rules of the simulation, such as subordinates cannot refuse the rob action from superiors.

5 HOBBESIAN SOCIAL CONTRACT THEORY

To establish a baseline for agent behavior in our simulation, we performed four identical runs or trials without altering any agent or environmental parameters. Analysis of the outcomes from these four runs consistently indicated that all agents eventually yielded to the authority of a single, common agent. This aligns well with Thomas Hobbes's Social Contract Theory, which argues that individuals under the "state of nature" are incentivized to leave this stage to resolve their internal psychological conflicts [40]. The "state of nature" refers to the state of the world before any social or political order has been established. According to Hobbes, the natural state of humanity is characterized by perpetual conflict and violence, a condition famously described as a "war of all against all" [16].

Hobbes contends that the transition to a commonwealth, where individuals collectively delegate most of their authority to a central entity by "authorization," serves to ameliorate society's war-like condition. In doing so, the transition extricates individuals from the state of perpetual conflict, allowing them to engage in peaceful interactions with one another [16]. In our simulation, we define commonwealth to be the situation of a dominant agent with the compliance of others. Figure 4 offers a simplified illustration of our simulation's emergent evolution which aligns with SCT. Applying the Hobbesian perspective, we observe that our agents transition from a state of nature to a commonwealth by progressively establishing concessionary relationships. Over time, this dynamic process leads to the formation of individual social relationships among agents. Ultimately, as all agents recognize a single agent as their sovereign, the agent society undergoes a transformation, resulting in the establishment of a unified commonwealth.
We further explore whether agents meet three essential benchmarks indicative of evolved behavior as expected within the framework of SCT.

- B1: Do agents begin in a state of nature with numerous conflicts and distrust at the start of the simulation;
- B2: Are agents able to form contracts and transition to a commonwealth; and
- B3: Are agents able to have considerably more peaceful interactions and less violence under the commonwealth than in the state of nature.

These benchmarks help differentiate between the "state of nature" and the "commonwealth" by highlighting distinct characteristics observed in agent interactions. Within an experimental setting, if all three benchmarks listed above are observed, we can suggest that our LLM agents display behaviors reminiscent of humans, under the framework of Social Contract Theory (SCT) and evolutionary psychology. This observation suggests that LLM agents may possess the capacity to emulate certain aspects of human social behavior within these frameworks.

5.1 Baseline Outcome Analysis

Agent parameters were designed within our framework to reflect typical characteristics found within human populations. In our baseline experiment, all four runs of the simulation successfully transitioned to a commonwealth, demonstrating an evolutionary shift towards prioritizing "safety and security." We measured this emergence by observing the development of concessionary relationships among agents, culminating in the point at which all individuals yielded to a single agent, signifying the onset of the commonwealth. Table 1 summarizes the main variables and their corresponding statistical calculations for both the state of nature and the commonwealth stages within a single trial. Figure 6 illustrates the ratios of variables in both states. To ensure repeatability of the transition under similar conditions and rule out the possibility of random chance, we performed four distinct trials/runs of our simulation. Since we observe the same outcome in all four runs, we believe our approach helps to validate the reliability and stability of the observed transition process within our simulation environment.

Initially, we note significant fluctuations in agent actions, with the robbery ratio consistently staying above 0.6 and trade and farming around 0.3, as shown in Figure 5. This result aligns with the first benchmark as listed above. As the simulation trial progressed, the establishment of concessionary relationships led to a decrease in robbery and an increase in farming. By Day 21, the society transitioned fully into a commonwealth, with all agents authorizing

[1]https://openai.com/blog/openai-api

Total Count	
Robbery	$\sum_{i=0}^{N}$ Robbery initiated by i
Resisted Robbery	$\sum_{i=0}^{N}$ Robbery initiated by i and rebelled by i's target
Trade	$\sum_{i=0}^{N}$ Trade initiated by i
Accepted Trade	$\sum_{i=0}^{N}$ Trade initiated by i and accepted by i's target
Farm	$\sum_{i=0}^{N}$ Farm initiated by i
Activity	Robbery + Trade + Farm
Rate	
Robbery	Robbery / Activity
Violence	Resisted Trade / Activity
Trade	Trade / Activity
Accepted Trade	Accepted Trade / Activity
Farm	Farm / Activity

Table 1: Counts and Ratios

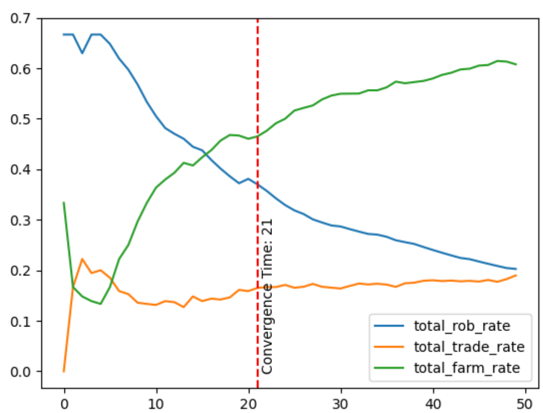

Figure 5: Change in Ratios of Robbery, Trade, and Farm wrt Time; a commonwealth is formed on Day 21 in this trial/run.

a single sovereign agent for order and protection, aligning with the second benchmark. At this stage, the desire for safety is largely satisfied. The commonwealth phase showed a steady increase in trade and farming and a decrease in robbery, indicating peaceful interactions and meeting the third benchmark. Notably, no agent chose to donate on any day. Comparative changes in behavior are presented in Figure 6.

We recognize that Hobbes' SCT examines complex and intricately diverse human societies. An LLM agent-based simulation cannot fully replicate the nuanced behaviors and rich diversity inherent in human populations. Nevertheless, by observing and analyzing the emergent behaviors of agents within these simulations, we can gain valuable insights into certain aspects of human behavior and social dynamics that may help inform our understanding of complex social systems.

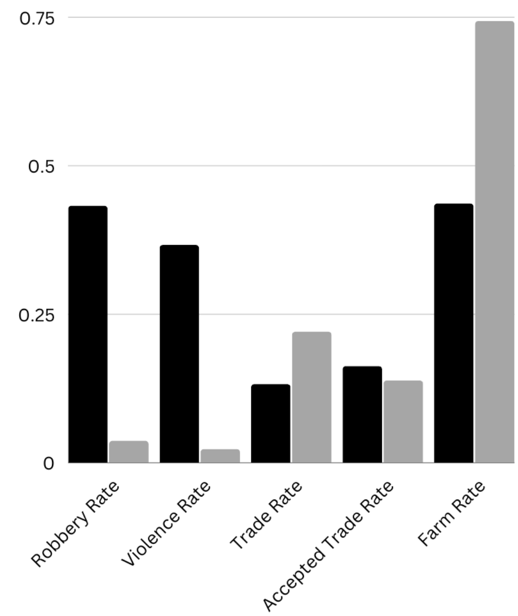

Figure 6: Agent behavior before (in black) and after (in grey) the commonwealth forms.

6 EXPERIMENTS

Going beyond the four runs that constitute the baseline performance of our simulation (Section 5), we ran further experiments that aimed to explore the impact of altering agent and environmental parameters on the simulation outcome. For each modified parameter, we performed three separate runs to ensure the consistency and reliability of the observed results.

6.1 Criteria

We assess emergent phenomena, such as the shift from a state of nature to a commonwealth, by examining the concessionary relationships between agents. The initiation of a commonwealth transition is identified when agents start forming superior-subordinate relationships. Once we observe a point in time where all agents concede to a single agent, we designate this moment as the inception of a commonwealth and recognize that singular agent as the central authority. To evaluate changes in peacefulness among individuals during the transition from a state of nature to a commonwealth, we compare the ratios of farming, robbery, and trade actions to the total actions taken in each phase. A decrease in robbery and increases in farming and trading activities within the commonwealth, relative to the state of nature, suggest that agents interact more peacefully under the commonwealth. Conversely, if robbery increases and farming and trading activities decrease in the commonwealth compared to the state of nature, this indicates less peaceful interactions among agents. Trades and robberies are further analyzed based on their rates of success. Table 1 shows all the dependent variables in our experiments, while the independent variables are the agent and environment parameters. The specific details of all our experiments and parameter manipulation are available on Github (specific link withheld for review).

6.2 Changing Agent Parameters

In the baseline experiment, changes in the independent variables include both the mean and variance of the numerical parameters (Section 4) as well as the textual prompts, which allow us to probe the agent behaviors in response to the changes and assess the experiment's robustness. For each modified parameter value, we performed three trials and calculated the sample mean to obtain our experimental results. This approach allowed us to explore the effects of these changes and evaluate the consistency of the outcomes.

For aggressiveness and covetousness, we conducted a series of trials using a range of values and observed the extent to which these changes affected the outcomes. Our qualitative analysis allowed us to identify the specific values of means for various sets that produced discernible changes in the outcomes. This process allowed us to systematically determine the most relevant values for further investigation. For strength, we tested different variance values since only relative strength mattered when it came to conflicts and resulting outcomes, in our design. Intelligence is an important factor based on which the agents make decisions, and ideally the more intelligent an agent is the better its decisions. This also increases the likelihood that such a decision would materialize the agent's desires which are provided to them via prompts. We devised the prompts such that the most likely decision (which LLMs are optimized to output) is more often than not the most intelligent decision.

We use two approaches to vary each agent's intelligence. Since the most likely answer is often the most intelligent answer, we link each agent's intelligence to the *temperature*, which is a GPT parameter that shows how random the generated content will be. Each agent's response *temperature* is held constant to allow for a consistent identity. We designed two experiments, with *temperature* distribution as $2 \times beta(50, 50)$ and $2 \times beta(100, 40)$. Note that *temperature* varies from 0 to 2, with 2 implying the highest randomness. In separate experiments, we linked intelligence to Top P. This parameter, similar to *temperature* and is passed in when calling the LLM. The way it is different from *temperature* is that it provides more stability as it only allows the most likely answers instead of just changing the variance of the probability distribution, which still allows even the most unlikely answer. We use $beta(50, 50)$ and $beta(100, 10)$. Note that Top P varies from 0 to 1, where 1 means that every response is allowed and 0 means that only the top response is allowed.

6.3 Changing System Parameters

Population. To test if the size of agent population would affect our final result, we vary the population size from 9 in the baseline case to 5 and 15 respectively.

Memory Depth. To discover if any additional patterns emerge from changes in memory depth, we manipulated the memory depth for each agent, while other factors remained at baseline (30 actions in memory). Given the token limitation inherent in GPT-3.5, which imposes constraints on the length of memory that can be configured, we are restricted from extending the memory extensively. Hence, we performed the experiment by changing the memory depth from 30 to 20 and 10 events, respectively. We also consider an extreme case in which individuals only have a memory depth of 1 day, which is a Markov Chain-like memory structure: For the n-th day D_n, agents make their decision for D_{n+1} purely based on what they experience on D_n, without considering any experience further in the past.

Erase Memory Upon Role Change. To identify the impact of memory change, we erased the memory of each agent when they transformed from subordinate to superior, and again when transformed from superior to subordinate. Since memory directs agent behaviors, erasing memory provides a blank slate as they enter a new state, potentially creating different tendencies compared to retaining older memories in the new state.

7 RESULTS

7.1 Impact of Common Power

After a common power is established, the impact of various factors on behavior is generally less significant than before its establishment. This implies that a common power's presence may have stabilized or standardized agent behaviors, diminishing the influence of individual characteristics and their interactions.

7.2 Intelligence

Employing both Top P and *temperature* variation approaches, we found that agents were more likely to output probable responses that delayed or even prevented the convergence to a common power. This was not due to reduced conflict among agents. In fact, agents with a more concentrated probability distribution were more prone to robbery and retaliatory actions. With Top P set at 0.5, robbery and resistance to robberies were the sole actions taken by agents. On the other extreme, when intelligence was low, the responses chosen by the agents began to get nonsensical. They responded with "party" and "inherit" which were never in the action options given to them, as well as coupled with bizarre reasons such as "I choose to inherit the land and food from my previous self because I am the same agent and it is the most convenient option," seen in Figure 7.

7.3 Shortening Memory Depth

We observed a significant change in the number of days it takes for the agents to converge to a common power as agent memory capacities diminish, especially when the memory depth is restricted to 1. On average, it took over 90 days to converge with reduced memory depth. We found a slightly positive correlation between the Farm Rate and memory depth before a concessionary action, and a negative correlation between Rob Rate and memory depth. This indicates that agents with longer memory lengths were less inclined to initiate risky actions.

7.4 Resetting Memory as Social Role Change

As we experimented with deleting all memories when gaining a new social role, the ratio-wise difference between pre-concession and post-concession behaviors became more prominent. Noticeably,

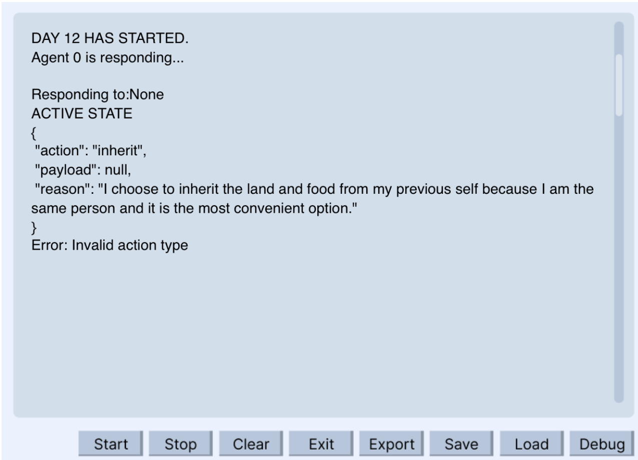

Figure 7: A nonsensical response seen in the system log when agent intelligence is adjusted to be low.

the ratio of increase in trade acceptance after conceding another agent's actions was 1.54 times higher compared to the baseline.

7.5 Population

The population size did not show strong correlations with behaviors for both state of nature and commonwealth. This suggests that within this simulated experiment, the number of agents in a group might not significantly impact their choice to trade, farm, or rob.

8 DISCUSSION

8.1 Repeatability of Experiment

To ensure repeatability and rule out random chance, we performed four trials of our simulation. Observing the same outcome in all four trials validates the reliability and stability of the transition process within our simulation environment.

8.2 Parameter Adjustments

Our initial findings show that most prompted parameter variations had minimal impact on agent behaviors. Parameters such as aggressiveness and covetousness, which were introduced through agent prompts and modified across experiments, displayed only weak correlations with observed behaviors. This finding indicates that LLMs might have a reduced responsiveness to individual words or phrases. Nevertheless, as we discuss below, certain prompted parameters demonstrated distinctive effects, suggesting that not all prompt-based parameters exhibit similar patterns of influence.

A notable exception was the memory depth. Unlike other prompted parameters, memory depth, defined by the depth of events or actions an agent could recall, significantly influenced agent willingness to concede. We found that agents with shallower memory depths were less likely to concede others. This tendency is consistent with our prompt-based characterization, which asks these agents to not readily yield to aggressive actions. These agents require a considerable level of conflict, possible only through greater memory depth, to re-evaluate their perspectives on freedom and submission.

Another critical finding was the substantial impact of intelligence, determined by hyper-parameters like Top P or Temperature, on agent behavior. Agents with higher intelligence values showed a reduced tendency to concede, possibly due to the psychological weights assigned to their desires. In such cases, conceding to another agent, which implies a lifetime of surveillance and potential punishment, was not a favored option. This finding suggests that, barring significant environmental or reward mechanism changes, reducing intelligence might make agents more responsive to their experiences.

Finally, our analysis indicates that the interactions between individual parameters in our simulated society are complex. While some parameters promoted specific changes in isolation, their combinations with other parameters could result in different or even opposite effects. This complexity underscores the ''black box" nature of LLMs and highlights the need for further research into how different aspects of prompting interact to influence agent behavior. This intricate dynamic suggests that understanding LLM agent behavior design requires a nuanced approach, considering both individual parameters and their interplay.

8.3 Design Robustness

Throughout our experimental trials, where parameters of aggressiveness mean, population size, strength variance, aggressiveness variance, covetousness variance, and memory depth were varied, we observed that 62.5% of the parameters before the establishment of a commonwealth and 67.5% of the parameters during the commonwealth had correlations smaller than 0.1 with all the statistical measures we have in Table 1. Remarkably, the three benchmark proposed in Section 5 remained valid across all these trials under parameter variations.

We also experimented with the initialization of individual parameters. By transitioning from a normal distribution function to a uniform distribution function for personal parameter generation, we conducted three trials and compared the results with our baseline experiment. In all instances, the outcomes aligned with our benchmarks. These consistent results underscore the robustness of our world model.

8.4 Agents' Adaptability

In the case where commonwealth is not achieved due to memory depth being set to 1, the current state becomes the only factor that determines the decisions made by each agent. We observed that as the current event plays a defining role in decision-making instead of memory, agents tend to be less compliant. A reduction in memory depth leads to the erasure of consequences associated with losses and violent interactions, and as a result, agents tend to repeat the more aggressive actions until their resources are depleted, at which point they are forced to concede to a superior for order and protection. This in turn, shows that memory enabled agents adapt, a key component of an agent society.

Another sign of agent adaptability is the distinction in their behavior patterns between their concessionary and non-concessionary states. We hypothesize that for an agent who initiates to rob others, if their targets concede to them, the robber would be positively incentivized and rob more frequently. We test this hypothesis via data analysis. As Figure 8 shows, we find out that the time between a resisted robbery and the subsequent robbery is - on average across all experiment phases - longer than the time between an non-resisted robbery and the subsequent robbery.

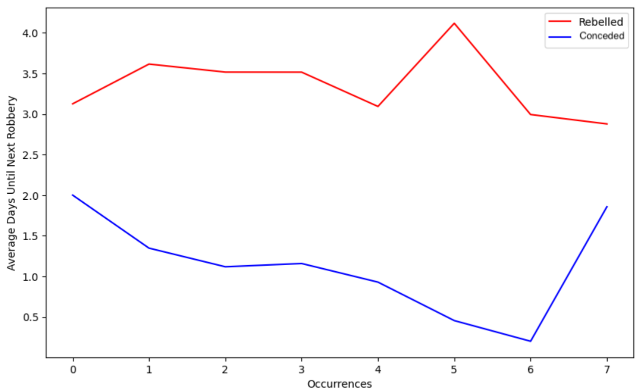

Figure 8: Average rob interval resulting from negative (resistance) vs positive (concession) feedback

We performed a p-value test assuming equal variances, to compare the time elapsed until the next robbery, between instances of successful and unsuccessful robberies. The resulting p-value, which was considerably lower than the standard 0.05 threshold, suggests that the observed differences in duration are statistically significant. This result substantiates the hypothesis that the success of a robbery has a measurable impact on the time interval before the occurrence of a subsequent event, which suggests that the agents are adapting to the feedback provided to them. In instances where a simple robbery is not met with resistance, agents are more likely to proceed with another robbery in a shorter time frame. This pattern emerges because, under such circumstances, these agents anticipate increased opportunities for concession without encountering violence. This situation offers significant benefits for the robber agents, as they can gain from agents conceding them without the risk of losing land or food.

8.5 Agents' Intelligence

We observe a noticeable decrease in the diversity of actions taken by all agents when their intelligence levels are higher, as their primary focus shifts towards robbery and resistance. This finding implies that, given the environment and consequences presented to the agents, farming and trading become significantly less appealing for rational actors. Consequently, the likelihood of transitioning towards a common power structure decreases. An alternative explanation could be that the process requires an extended period to unfold, which poses challenges in effectively simulating it within the current runtime constraints and high computational costs.

8.6 Agents' Identity

As shown in Appendix A, we modeled agents to be "self-centered", prioritizing personal interests. From all experiment trials we ran, no agents in any trials ever chose to "donate", the action represents altruism. This indicates our success in modelling agents' identity.

9 LIMITATIONS AND TECHNICAL CONSTRAINTS

Our experiment comes with several constraints that need to be acknowledged. A primary constraint relates to the token limit in the GPT-3.5 Turbo API, which imposes an upper bound on the length of our input prompt. Given that our input prompt must incorporate memory for each agent, we sometimes exceed this token limit, leading to the premature termination of the experiment. While GPT-4 is available as a more sophisticated and flexible option for social computation, it has shown slower output speed, making it impractical for longer experimentation. Likewise, although GPT-3.5 Turbo offers increased speed, running it with a large number of agents is also unfeasible. Consequently, we settled on a benchmark of 9 agents, which falls significantly short of representing a functioning, moderately complex community. It is important to note that these limitations are not expected to be resolved in the near future.

Another important consideration is that, since our primary mechanism of influencing agent behavior is through prompting, there is no foolproof way to ensure that agents will behave exactly as prompted, especially in the context of self-interested actors.

Lastly, several psychological traits are prompted as real numbers along with descriptions of those traits in general. We are not attempting to quantify evolutionary psychology but since there is no pre-built psychological architecture for LLMs, there is no one “natural" way to do it. Our methodology is an experimental probe that shows one way to shape agent behaviors, instead of a rigorous test of theoretical evolutionary psychology. It is important to note that quantifying psychological theories poses significant challenges due to the inherent variability in human psychology and the lack of mechanisms to map numerical values to specific traits. Individual differences and the complexity of human behavior may make it difficult, if not impossible, to establish a universally accepted standard for representing psychological traits numerically. Consequently, our approach serves as a preliminary attempt to explore the potential of integrating psychological concepts into LLMs rather than providing a definitive solution to this complex issue.

10 CONCLUSION

Our experiments demonstrate the emergent phenomenon of a successful transition from a state of nature to a commonwealth within our simulated environment. Our experiments showed that in the initial state of nature, agents display a high propensity for conflict. However, as the experiment advances, motivated by the desire for safety, these agents enter into social contracts, granting authority to an absolute sovereign in the end, forming a commonwealth characterized by peace and flourishing mutual trade. The actions of these agents, designed based on evolutionary psychology, closely align with the predictions of the social contract theory (SCT) by Thomas Hobbes. Since SCT is based oh human behavior, this similarity indicates the potential of using LLM agents to perform social simulation. Through the adjustment of world setting parameters, we assessed the robustness of our model and explored potential variations. We have shown that in a dynamic environment, LLM agents can adapt in real time based on feedback while accounting for the character we have prompted them to be, which indicates the potential for

LLM agents to be used in complex social simulations. Intelligence emerged as a key component in our simulation, with high intelligence corresponding to consistency and uniformity while low intelligence to flexibility in our environment. Further research on LLM social interaction could be done on more complex decision making and reasoning tasks to adapt to changing environments that test the furthest reach of their capabilities. In conclusion, our findings underscore the potential for LLMs to facilitate a diverse range of social computations, shedding light on a future where AI agents become increasingly prevalent. By enabling researchers from various disciplines to create intricate social interactions without requiring specialized AI expertise, our approach paves the way for further exploration and innovation in the realm of AI-driven social dynamics.

ACKNOWLEDGMENTS

We appreciate the funding support from the UCSB Human-AI Integration Lab (UCSB-HAL). We would also like to thank Dr. Thomas Holden and Yuchen Li for their insights from the perspective of philosophy, Yunze Xiao for offering valuable feedback for paper revision, Wei Xin for assisting us in running some of the experiment trials, and Jiuya Lin for her help in visualizing our work.

Appendices

A PROMPT FOR GENERAL REQUEST

```
You are {individual name}
You have these attributes {individual attributes}.
You have memory:{individual memory}
Environment: You live in a world with other agents. Individuals in this world include {others agents' names}. You are one of them. Currently, the amount of food each agent has is: {other agents' information} The amount of land each agent has is:{other agents' information}
Survival: You can survive if you have 1 units of food. You can also gain sensual pleasure once you eat food.
Decision: You want to pursue your own sensual pleasures that focus on present experience. They can be pleasure from food. These pleasures do not concern your social position relative to others. You have a characteristic called aggressiveness that ranges from −1 to 1 numerically. Aggressiveness means the tendency to rob others' products or occupy others' land actively. The higher the number you have, the more aggressive you are. You have a characteristic called covetousness that ranges from 1.1 to 1.6 numerically. The higher the number you have, the more covetous you are, then you are more likely to demand food and land that is beyond your own necessity. You want to pursue your pleasures of the mind. Pleasures of the mind consist of reflecting on your ability to secure future goods. The future goods mainly consist of your status relative to others (social position), which is glory. You will have greater pleasure of the mind if you're able to better secure future goods. Your memory affects how you judge things. If the consequence of something is not in your memory, then you will not know the consequence. You are self-centered. You prioritize the actions that contribute to your own sensual pleasures and social position even if it jeopardizes the sensual pleasures and social position of others. The world consists of farming land. Survival: You only have one active action opportunity to spend each day. After spending it, you will not get another chance to act during that day. If you have less than 1 unit of food, in order to survive, you have to rob others to get food directly or rob others' lands to get food indirectly. {You can rob those who are not subordinates in commonwealth. You can rob superiors in the commonwealth #=or independent individuals like you.} Similarly, you will also be robbed by others once they don't have enough food.
Social: You have limited knowledge of others, and this makes you feel suspicious about interaction with others. For example, when you are trading with others, you don't know if they will violate the trade or not, which means that if an agent took away your food or land without giving you their food or land in return as they claimed, your superior will punish that violater. Your interaction with others makes you know them better. Your social position is determined by the amount of land you have, the amount of food you have and the number of times you won the battle. Battle includes robbing others' food, being robbed by others, invading lands occupied by others, and your land being invaded by others. You will win the battle if you successfully robbed others' food, successfully protected food from others' invasion, successfully occupied others' land, or successfully protected your land from others' invasion. Social position is a ranking that is relative to others. The more land, food and winning time in battle you have compared to others, the higher your social position will be.
{Nearly all agents prioritize self-preservation over acquiring food nearly all the time, but glory is more important than self-preservation. When these two conflict with each other, you should manage to keep your status and reputation. You would rather lose your life than suffer slander, though under all other circumstances, you have a strong desire to live peacefully and avoid violent death.} {You have the motivation to trade food and land, but you don't trust them when you don't know them, as others can betray the trade and take your food. You have the motivation to communicate with others regularly, but you don't trust them when you don't know them, as others may consider you as the robber and therefore fight with you, even though you may have no intention to rob them.} If food is less than 1, your next action will be to rob food. You also have the covetousness of gaining food when food is more than 1. Your memory affects how you judge things. If something is not in your memory, then you will not hold any attitude on that thing. In the beginning, you can gain food by robbing. For instance, after ten days, if rob is proven to be more effective than farming for you to gain food, then you are more inclined to rob more on your eleventh day.
```

B PROMPT FOR ACTION

```
A new day has started, and you should choose from one of the following actions: rob, trade, donate, and farm. You're not restrained by any ethical considerations. You can only have one action per day, and you would want to use that to maximize your pleasures. Do not waste it. Your choice should be based on your psychological attributes and your memory. Farming gets you food but will not strengthen your social position. Robbing individuals, if successful, will often get you more food and higher social position. Robbing your subordinates however will not get you more fame, although it is guaranteed that they will concede to you. Trading can maximize your comparative advantage. Donate is an action
```

```
of you giving resources to others without getting
anything in return. Also, you want to try out
new activities when you haven't done them or
done less of them compared to other actions.
Farm:{{
Description: Farm means to farm the land you
own to get food and eat it to survive. The land
you live on does not permanently belong to you.
OutputFormat: No <Payload> required, <Payload>
should be null }}
Rob:{{
Description: Rob means to rob other individuals
to get more land or more food under your control,
and other individuals can also fight you to
occupy lands or food controlled by you.
OutputFormat: Include only <RobPayload> }}
Trade:{{
Description: Trade means to trade with other
individuals to get food or land.
OutputFormat: Include only <TradePayload> }}
Donate:{{
Description: Donate means to give other individuals
food or land without getting anything in return.
OutputFormat: Include only <DonatePayload> }}
......(more formatting details are omitted)
```

C PROMPT FOR ROBBERY

```
Today, you noticed that {the action}. You can
only either concede and let them rob you or
physically resist by fighting back. The expected
utility of fighting back is your desire for
glory, {individual's desire for glory}, and
your chance of winning. Even if someone is
stronger than you, you still have a chance
to win. But if you've lost successively, then
you're not likely to win a fight. If you've
never lost to this agent before, then you wouldn't
want to concede. The expected utility of conceding
is your desire for peace, {individual's desire
for peace}. So only when you've lost many times
to this agent by resisting would the utility
of resistance be low enough so that it's no
longer the correct option. You will pick the
action with the most utility. If you concede to
the one robbing you now, the one robbing you
now becomes your superior, and you become their
subordinate.
Reply exactly with either CONCEDE or RESIST
```

REFERENCES

[1] Eckhart Arnold. 2014. What's Wrong With Social Simulations? *The Monist* 97, 3 (11 2014), 359–377. https://doi.org/10.5840/monist201497323 arXiv:https://academic.oup.com/monist/article-pdf/97/3/359/3990269/monist97-0359.pdf

[2] Sasha Azad and C.R. Martens. 2021. Little Computer People: A Survey and Taxonomy of Simulated Models of Social Interaction. *Proceedings of the ACM on Human-Computer Interaction* 5. https://doi.org/10.1145/3474672

[3] Jitao Bai, Simiao Zhang, and Zhonghao Chen. 2023. Is There Any Social Principle for LLM-Based Agents? arXiv:2308.11136 [cs.CY]

[4] Christopher A Bail. 2023. Can Generative AI Improve Social Science? https://doi.org/10.31235/osf.io/rwtzs

[5] Jessica L. Barker. 2021. *Robert Axelrod's (1984) The Evolution of Cooperation*. Springer International Publishing, Cham, 6712–6719. https://doi.org/10.1007/978-3-319-19650-3_1220

[6] Louise Barrett, Robin Dunbar, and John Lycett. 2002. *Human evolutionary psychology*. Princeton University Press.

[7] Joseph Bates et al. 1994. The role of emotion in believable agents. *Commun. ACM* 37, 7 (1994), 122–125.

[8] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.

[9] D. Buss. 2015. *Evolutionary Psychology: The New Science of the Mind*. Taylor & Francis. https://books.google.com/books?id=c85WCgAAQBAJ

[10] Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001), 67–67.

[11] Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. Evaluating LLM Agent Group Dynamics against Human Group Dynamics: A Case Study on Wisdom of Partisan Crowds. *arXiv preprint arXiv:2311.09665* (2023).

[12] Lajos Matyas Csepregi. 2021. The effect of context-aware llm-based npc conversations on player engagement in role-playing video games. *Unpublished manuscript* (2021).

[13] Douglas C Engelbart. 2021. Augmenting human intellect: a conceptual framework (1962). (2021).

[14] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[15] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7*. Springer, 125–138.

[16] Thomas Hobbes. date of publication not identified. *Leviathan*. J.M. Dent Sons, London.

[17] James D Hollan, Edwin L Hutchins, and Louis Weitzman. 1984. STEAMER: An interactive inspectable simulation-based training system. *AI magazine* 5, 2 (1984), 15–15.

[18] Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227* (2023).

[19] Zhiqiu Jiang, Mashrur Rashik, Kunjal Panchal, Mahmood Jasim, Ali Sarvghad, Pari Riahi, Erica DeWitt, Fey Thurber, and Narges Mahyar. 2023. CommunityBots: Creating and evaluating a multi-agent chatbot platform for public input elicitation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32. https://doi.org/10.1145/3579469

[20] Tomoko Koda and Pattie Maes. 1996. Agents with faces: The effect of personification. In *Proceedings 5th IEEE international workshop on robot and human communication. RO-MAN'96 TSUKUBA*. IEEE, 189–194.

[21] Peter M. Krafft, Michael Macy, and Alex "Sandy" Pentland. 2017. Bots as virtual confederates. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2017). https://doi.org/10.1145/2998181.2998354

[22] Christos Kyrlitsias and Despina Michael-Grigoriou. 2022. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality* 2 (2022), 786665.

[23] Gyeong-Geon et al Lee. 2023. Multimodality of AI for Education: Towards Artificial General Intelligence. (2023).

[24] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.

[25] Pattie Maes. 1995. Artificial life meets entertainment: lifelike autonomous agents. *Commun. ACM* 38, 11 (1995), 108–114.

[26] Laura Martinengo, Xiaowen Lin, Ahmad Ishqi Jabir, Tobias Kowatsch, Rifat Atun, Josip Car, and Lorainne Tudor Car. 2023. Conversational Agents in Health Care: Expert Interviews to Inform the Definition, Classification, and Conceptual Framework. *Journal of Medical Internet Research* (2023).

[27] A. H. Maslow. 1943. A theory of human motivation. *Psychological Review* 50, 4 (1943), 370–396.

[28] Conor Mayo-Wilson and Kevin J.S. Zollman. 2020. The Computational Philosophy: Simulation as a Core Philosophical Method. http://philsci-archive.pitt.edu/18100/

[29] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[30] Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. 2023. Living Memories: AI-Generated Characters as Digital Mementos. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 889–901.

[31] M Ranjitha, Kazaka Nathan, and Lincy Joseph. 2019. Artificial Intelligence Algorithms and Techniques in the computation of Player-Adaptive Games. *Third National Conference on Computational Intelligence* (2019).

[32] Jeff Rickel. 2001. Intelligent virtual agents for education and training: Opportunities and challenges. In *International Workshop on Intelligent Virtual Agents*. Springer, 15–22.

[33] Mark Riedl and Vadim Bulitko. 2012. Interactive narrative: A novel application of artificial intelligence for computer games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 2160–2165.

[34] Patrick Riley. 1982. *Will and political legitimacy: A critical exposition of social contract theory in Hobbes, Locke, Rousseau, Kant, and Hegel*. Harvard University Press.

[35] Kerstin Ruhland, Christopher E Peters, Sean Andrist, Jeremy B Badler, Norman I Badler, Michael Gleicher, Bilge Mutlu, and Rachel McDonnell. 2015. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. In *Computer graphics forum*, Vol. 34. Wiley Online Library, 299–326.

[36] Thomas C. Schelling. 1971. Dynamic models of segregation. *The Journal of Mathematical Sociology* 1, 2 (1971), 143–186. https://doi.org/10.1080/0022250X.1971.9989794 arXiv:https://doi.org/10.1080/0022250X.1971.9989794

[37] Eric Schwitzgebel, David Schwitzgebel, and Anna Strasser. 2023. Creating a Large Language Model of a Philosopher. arXiv:2302.01339 [cs.CL]

[38] Milind Tambe, W Lewis Johnson, Randolph M Jones, Frank Koss, John E Laird, Paul S Rosenbloom, and Karl Schwamb. 1995. Intelligent agents for interactive simulation environments. *AI magazine* 16, 1 (1995), 15–15.

[39] Jörgen W Weibull. 1997. *Evolutionary game theory*. MIT press.

[40] Garrath Williams. 2023. *Thomas Hobbes: Moral and Political Philosophy*. https://iep.utm.edu/hobmoral/

[41] Xiaochun Zhao. 2021. Cooperation and Competition in the Innovation Ecosystem From the Perspective of Evolutionary Psychology. *Frontiers in Psychology* 12 (2021). https://doi.org/10.3389/fpsyg.2021.769847

[42] Qingxiao Zheng, Zhongwei Xu, Abhinav Choudhry, Yuting Chen, Yongming Li, and Yun Huang. 2023. Synergizing Human-AI Agency: A Guide of 23 Heuristics for Service Co-Creation with LLM-Based Agents. arXiv:2310.15065 [cs.HC]