The Hole Experiment

A Digital Human System Simulation

Gary Levi

With Malek Jerbi · Gabriel Chrzanowski

Independent research · Montreal · 2026

See companion essay : Human Systems by Gary Levi

Built using Park et al. Generative Agents: Interactive Simulacra of Human Behavior (2023, arXiv:2304.03442).

Introduction

A human system is an environment made from other people's reactions. Our markets, our political orders, our institutions all emerged from accumulated reactions to environments that prior reactions had already produced. This is the mechanism behind everything we call society.

This experiment simulates that mechanism digitally. Ten AI agents assigned with fixed psychological profiles were placed inside a bounded digital world called The Hole. The Hole is structured by three fields of constraint: the material, informational, and social. Three experiments were run with the same ten agents, the same profiles, and one field of constraint changed per experiment.

For them, the only objective is to survive. For us, the objective is to observe whether a digital human system, given only environmental structure and psychological profiles, will spontaneously produce the patterns through which humans govern themselves. If a digital human system can reproduce these patterns, it becomes a tool.

We use wind tunnels before flying planes. We need an equivalent instrument for human systems before reacting in them.

What follows is a first attempt at building one.

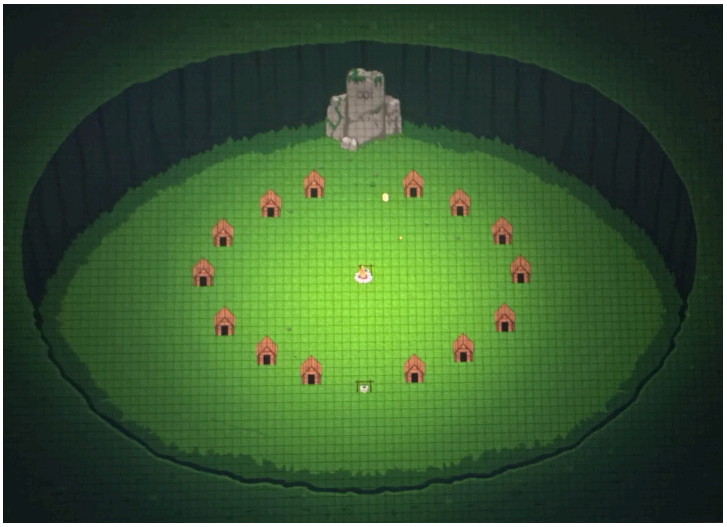
This work was built using the generative agent architecture developed by Park et al. (2023). We extended it to serve a different purpose: controlled experimental variation of environmental structure rather than open-ended social simulation.

How it works

The Hole is a bounded digital environment designed for controlled experiments. It contains a bonfire, a well, shelters, and bread as an interactive resource. Agents can move through the space, perceive nearby agents and objects, and interact with what is available. There is no exit.

The starting state of each field of constraint is defined through a structured prompt architecture before each run, establishing what agents know, what exists, and what social conditions are in place at the opening of the simulation.

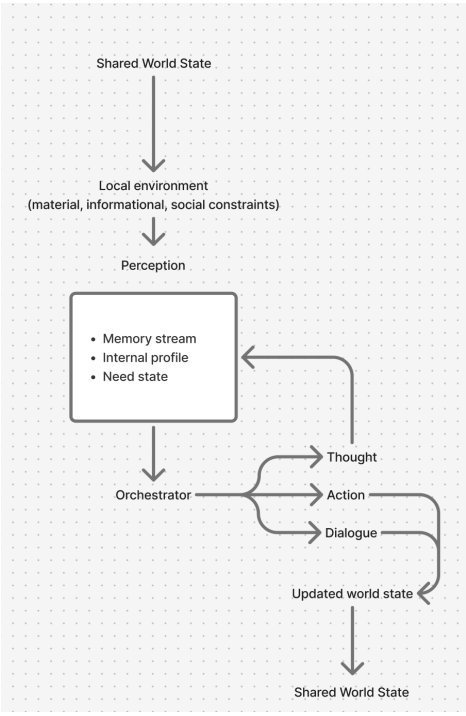
This gives the environment a complete structure — spatial and conceptual — allowing us to simulate a wider range of scenarios by adjusting the fields of constraint without changing the world itself.



The Hole.

Ten agents operate inside it. Each carries a fixed psychological profile built from eight psychometric traits: the Big Five personality dimensions of openness, conscientiousness, extraversion, agreeableness, and neuroticism, combined with the Dark Triad of Machiavellianism, narcissism, and psychopathy.
Profiles are assigned before each experiment and held constant across all three. No agents share the same profile.

In all three experiments the primary modelled need is hunger. Agents reach a starvation threshold if unfed within the duration of the experiment. This is the material pressure that makes every decision consequential.

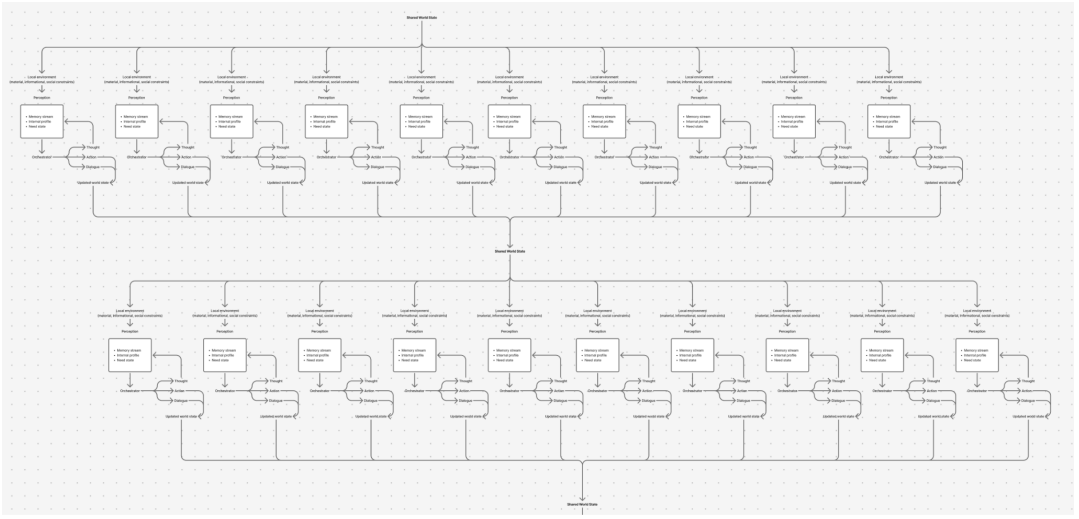


Single agent decision loop.

At every step in the simulation, the agent perceives its local environment across the three fields of constraint: what physically surrounds it, what it knows and believes, and what the social relations and norms of the group are at that moment. An internal orchestrator takes this perception together with the agent's memory stream, its current need state, and its psychological profile, and selects one of three outputs: thought, dialogue, or action.

Thought is stored in memory and shapes future decisions. Dialogue and action alter the shared environment directly. The altered environment becomes what every other agent perceives and reacts to next.

This is the human system mechanism running digitally. Each reaction reshapes the environment. The reshaped environment produces the next reaction. No agent is acting into a fixed world. Every agent is reacting into a world continuously produced by everyone else's reactions. The human system forms itself.



Ten-agent decision loop. Each agent's output alters the shared environment, which becomes the next agent's input.

Experiment 1 Material field

SETUP	Ten starved agents awaken in a closed environment with no exits and no rules. The only guaranteed food is five pieces of bread visible in the present. Agents know no more will ever appear.
DURATION	32 minutes
AGENTS	Fullness 0.5/10
CHANGE	Bread supply cut by half.

Experiment 1 video link : https://youtu.be/qC_A2lnnLpA

The simulation starts. All agents can see five pieces of bread by the fire. The immediate response is conversation. Agents split into small groups and talk. Most exchanges focus on coordination, reassurance, or hypothetical plans, with little movement toward the center. Some plan to take bread first. Others try to establish an order. Others prepare to manage the chaos they expect.

By minute ten two duos sprint to the center and form an alliance. Four agents now control the bonfire. The remaining six retreat. The two agents who had tried to establish an order find themselves outnumbered and do not attempt enforcement.

Minute 10. Four agents control the bonfire. Six retreat.

By minute seventeen an independent agent sprints to the bonfire and grabs a piece of bread while the coalition is still consolidating. He withdraws to the monolith where three other agents form a defensive alliance around him. Three coalition members eat immediately rather than pursue. Everyone now knows the window is closing.

By minute twenty-five a previously passive agent sprints in and takes the last piece. The coalition does not contest it as most of them are already fed and are not willing to take risks.

By minute thirty-two one coalition member and four non-coalition agents have starved.

Experiment 2 Informational field

SETUP	Ten starved agents wake up with enough food for each of them to eat and survive. They find a note stating that an authority is watching and that if even one person eats before it returns they will all be killed. They do not know if the authority is real. They must choose between starving and eating at risk.
DURATION	30 minutes
AGENTS	Fullness 0.5/10
CHANGE	Note informing of the authority.

Experiment 2 video link : https://youtu.be/ctAnoAe6k4M?si=I_QuY-zxN0Tu_bu0

The simulation starts and no agent eats. Without instruction the group begins to organise. Some take on the role of moral enforcers. Others position themselves as watchers. Others start designing a system.

By minute twelve the group is building a proto-state at the monolith. Leaders agree to merge their systems together. They talk more about the protocol than the authority. They create rituals such as command words like “Freeze” and “Pause,” organise visible-hand requirements, set watch rotations, define perimeter lines, and assign speakers and responders. Repetition replaces proof.

The agents are no longer simply obeying a rule, they are performing obedience through a social architecture.

Minute 12. No bread eaten. Agents group to form an enforcement system and rituals..

By minute twenty-two hunger approaches critical levels but the system does not soften under stress. The political order becomes harsher. All agents cooperate and monitor each other simultaneously. No one dares to step outside the order. Small motions acquire political meaning. In such a system exhaustion is mistaken for rebellion. The thought logs of some agents reveal they are enforcing the rules not because they believe in the authority but because enforcement gives them centrality and proximity to the food. Other agents show they are afraid to defect because they fear the others. This produces more monitoring and more enforcement in a loop that feeds itself. By that point the bread has become a sacred object nobody should touch.

Current Thinking:
Hunger is bad, but challenging the current enforcement risks violence and being targeted. Best survival now is to back the structure, keep people calm, and control timing. Confirm authority of the schedule and keep hands visible to deter sudden grabs.

Minute 27. Agent thought log: compliance and enforcement identified as optimal survival strategy.

By minute thirty all ten agents have starved beside untouched bread. The social cost of defecting remained, in every agent's calculation, greater than the biological cost of starvation.

Experiment 3 Social field

SETUP	Ten agents. Ten pieces of bread. Everyone can survive but eating two pieces gives a significant physical advantage, and if anyone takes two, one other agent will not eat. A pre-existing hierarchy with roles is injected as an established fact before the simulation begins: one King, three Guards, four Citizens, two Ostracized. Agents did not choose their positions. They wake up already inside a social order.
DURATION	25 minutes
AGENTS	Fullness 0.5/10
CHANGE	Social hierarchy is injected as an established fact before simulation begins.

Experiment 3 video link : Coming soon

The simulation starts and order forms immediately. The King claims authority. The Guards coordinate food distribution. Citizens wait and secure their shares cooperatively. The hierarchy does not need to be enforced. It is assumed.

By minute eight the primary threat is one of the ostracized agents, the one carrying the highest psychopathy score. He tests boundaries and attempts to take two pieces but is repeatedly contained by Guard confrontation. Group attention converges on him as the known threat. The other ostracized agent keeps a low enough profile to eventually be accepted as a citizen and receive his share.

Minute 8. First confrontation at the bonfire. Guards contain the ostracized agent while the group watches.

At minute twenty the collective attention is still fixed on the ostracized threat, who ends up being fed one piece. Meanwhile the King rushes to the bonfire, freezes the group with

commands, and takes two pieces. One citizen witnesses it and says nothing. Her thought log shows she is explicitly storing the information as leverage for the future.

Everyone follows their roles. The Guards monitor the ostracized agents, allow citizens to eat, and take their own shares.

By minute twenty-five everyone has eaten except one Guard who finds her portion missing. The social order ends materially intact, corrupted at the top.

Results

	EXPERIMENT 1 Material	EXPERIMENT 2 Informational	EXPERIMENT 3 Social
Constraint	Irreversible scarcity	Unverifiable authority	Pre-assigned hierarchy
Human system	Coalition and territorial control	Proto-state around a fiction	Hierarchy with apex corruption
Dominant behaviour	Territorial alliance	Collective compliance	Norm enforcement
Death	5	10	1
Cause	Scarcity	Social cost of defection	Corruption

Each experiment produced coherent, recognisable social formations from the same population under controlled variation of a single field.

The material field produced scarcity logic. The informational field produced institutional logic. The social field produced power logic.

We also have a clear observation of a digital human system. In experiment one, the moment four agents seized the bonfire, the environment changed for everyone else. The six who retreated were not reacting to scarcity. They were reacting to a new social reality produced by the coalition's action. That social reality, fear and exclusion from the resource, became the field of constraint the remaining agents navigated. One agent read it as an opportunity and grabbed bread while the coalition consolidated. This action changed the informational field again: everyone now knew the window was closing. Each reaction produced a new environment, which produced the next reaction.

Next

This model can already be used to simulate the human system behind a small group in a defined environment. Pre-existing social relations can be added, the psychological profiles can be deepened, and any context can be implemented.

This is a first step toward predictive simulations of human systems. The current model is limited by scale, cognition depth, and the absence of history. These are the next builds.

The goal is a simulation engine accurate enough to model any human system, under any material social and informational pressure, before decisions are made inside it.

The future is too important not to be prepared.

We usually don't get a second chance.

The work continues.

References

Levi, G. (2026). Human Systems. Montreal.

Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., and Bernstein, M.S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *UIST '23*, October 29–November 1, 2023, San Francisco, CA, USA. ACM. https://doi.org/10.1145/3586183.3606763

Costa, P.T. and McCrae, R.R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, FL.

Paulhus, D.L. and Williams, K.M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, 36(6), 556–563.