LLM Agent Driven UBI Policy Experiments: A Case Study of AgentSociety as AI Social Scientist Platform

Che Bai
Tsinghua University
Beijing, China

Yuwei Yan
The Hong Kong University of Science and Technology
Hong Kong, China

Jinghua Piao
Tsinghua University
Beijing, China

Yong Li
Tsinghua University
Beijing, China

Abstract

Universal basic income (UBI) experiments are informative but costly to implement at scale and difficult to extend to future shocks. This paper studies an alternative research workflow for UBI built on AgentSociety: human researchers work with AI social scientists to specify hypotheses and experiments, while LLM agents act as silicon subjects in the designed policy environment. First, we reproduce three classic settings—the Finnish basic income experiment, Kenya GiveDirectly, and the U.S. Stockton SEED program—and compare simulated outcomes with their real-experiment targets. Second, we trace policy dynamics through baseline, intervention, and withdrawal phases. Third, we extend the experiment to an AI labor-replacement shock with UBI and AI-tool interventions. The results show that AgentSociety closely tracks real-study benchmarks in terms of income, consumption, labor participation, and psychological security. Under AI shock, UBI reduces hardship but is insufficient to restore employment and earnings; combining UBI with AI tools and training yields the strongest adaptation response. The results support AgentSociety-based UBI experiments as a low-cost, repeatable digital policy laboratory for mechanism discovery and counterfactual exploration.

CCS Concepts: • **Human-centered computing; Social and professional topics**;

ACM Reference Format:
Che Bai, Yuwei Yan, Jinghua Piao, and Yong Li. 2026. LLM Agent Driven UBI Policy Experiments: A Case Study of AgentSociety as AI Social Scientist Platform. In . ACM, New York, NY, USA, 10 pages.

KDD'26, Jeju, Korea

1 Introduction

Universal basic income (UBI)a social welfare proposal where all population receive a minimum income with the form of an unconditional transfer payment, is an unusually demanding object of experimental study. Its effects are simultaneously economic and social: the transfers may alter consumption, debt management, hours of work, job search, family support, and longer-run investment. Although field experiments remain indispensable, they are structurally constrained by high prohibitive costs, protracted timelines, and different selection biases. A real experiment must choose a single target population, a specific transfer schedule, a unique institutional environment, and a restricted set of outcomes. This burden spans urban guaranteed-income demonstrations, universal dividends, and transfers in rural poverty settings [11, 13, 20]. As a result, the available evidence can identify important effects of realized experiment while leaving many policy variations untested.

The problem becomes more difficult when the research target is a future labor-market transition rather than an observed historical environment. Generative AI can raise productivity for some tasks and workers [8, 16], while the broader investigation makes displacement and wage pressure plausible concerns [1, 4]. A policymaker may therefore ask questions that cannot yet be answered by an existing randomized trial: Does unconditional cash merely cushion an AI-driven earnings loss? Does it release the time and psychological bandwidth required for retraining? Does cash become more effective when access to AI tools and training is bundled with it? Directly launching every such trial in the field before the shock is observed is neither timely nor economically realistic.

This paper studies an alternative complement to field experimentation: an LLM agent based social experiment that is calibrated against known UBI evidence before it is used for counterfactual exploration. In this situation, LLM agents are not human subjects, and a simulated treatment contrast is not a causal estimate for an actual population. However, if a specified agent architecture can reproduce salient directional

regularities from several external experiments, the same architecture may offer a useful, low-cost environment for identifying mechanisms, revealing implausible assumptions, and prioritizing subsequent field studies. This calibration-first stance elevates policy simulation from a form of unsupported prediction to a rigorous tool for scientific hypothesis generation.

We instantiate this idea in AgentSociety 2 with the AI social scientists workflow[18], which is designed to disentangle the dual roles often conflated in LLM-based social simulations. In this platform, *Silicon Scientists* support the human researcher in benchmark collection, hypothesis formulation, scenario design, and interpretation. *Silicon Subjects* are the agents exposed to treatment: low-income workers represented with household constraints, occupations, income and balance-sheet states, psychological burden, and memory of preceding monthly outcomes. The distinction matters: Silicon Scientists help make the experimental assumptions explicit; Silicon Subjects generate behavior within those assumptions and procedures. The human researcher retains responsibility for deciding what counts as external evidence and for interpreting the boundaries of any finding.

The design follows two stages. In the first phase, we construct replicas of three canonical cash-transfer experiments selected to cover diverse socio-economic settings: (i) the Finnish basic income experiment, representing a high-welfare unemployment context; (ii) GiveDirectly Kenya, capturing a rural, low-income development setting with productive-investment channels; and (iii) the U.S. Stockton SEED program, exemplifying a modern urban guaranteed-income demonstration. We examine whether the LLM-agent society tracks real-experiment benchmarks for income, consumption, labor participation, and psychological security, and whether policy introduction and withdrawal generate coherent temporal responses. In the second stage, we introduce an AI labor-replacement shock and compare control, UBI, AI shock, UBI + AI shock, and UBI + AI tools/training arms. This sequence is intentional: a novel counterfactual environment is analyzed only after the simulation has been validated against known regularities.

This paper makes the following three contributions. First, we formulate an AI social scientists mode of computational policy research in which researchers and AgentSociety iterate over benchmarks, hypotheses, experimental protocols, and conclusions. Second, we develop an LLM-agent-driven UBI experimental environment with constrained household states, monthly decisions, policy phases, and cross-context calibration. We find our classic replicas align with key outcomes from Finland, Kenya, and Stockton. Third, we extend the calibrated environment to an AI labor-replacement counterfactual. In the simulation, AI shock lowers employment and earnings and raises distress; UBI chiefly buffers hardship, whereas UBI combined with AI tools and training produces the strongest adaptive income trajectory.

2 Related Work

The Finnish basic income experiment studied unemployed benefit recipients in a high-welfare environment and documented improved security with limited employment changes [14]. GiveDirectly's Kenya programs study poor rural households and illuminate consumption, enterprise, saving, and payment-timing mechanisms [6, 9, 11]. Evidence on conditional transfers likewise demonstrates that policy rules alter behavioral channels beyond transfer amounts [5]. U.S. unconditional-cash evidence similarly reports effects on household spending, labor outcomes, and wellbeing [7, 13, 15, 19, 20]. We use the three experiment contexts selected in the study blueprint—Finland, GiveDirectly Kenya, and Stockton SEED—as cross-context validation targets.

Generative agents provide persistent profiles, memory, and action traces for interactive simulation [17]. LLM-generated samples and simulated economic agents provide early evidence that model responses can be used to replicate or stress-test human-subject patterns, while also making validity a central concern [2, 3, 12]. AgentSociety advances this direction by supporting large-scale LLM-driven social simulations and AI social-scientist workflows [18]; surveys discuss their opportunities and validity challenges [10]. Meanwhile, experiments with generative AI tools find productivity gains for knowledge work and customer support [8, 16], while automation research documents the possibility of labor displacement [1, 4]. These streams motivate our core counterfactual: whether UBI alone protects workers under AI shock, or if access to AI tools is necessary for adaptive recovery.

3 Method

Our method casts UBI investigation as a cooperative and auditable simulation workflow rather than a single prompt asking an LLM to forecast a treatment effect. A human researcher specifies the substantive question and candidate benchmarks. AgentSociety then supports four linked operations: grounding a scenario in empirical trial parameters, constructing heterogeneous Silicon Subjects within their monthly environments, executing treatment and counterfactual arms under shared measurement rules, and interpreting trajectories against external benchmarks. This sequential modularity effectively decouples errors in scenario construction from the intrinsic behavioral variances generated by simulated subjects.

The workflow contains two agent roles. *Silicon Scientists* are research-assistance agents: they organize prior findings, translate a policy idea into observables, define experimental contrasts, and summarize deviations from benchmarks. They never constitute treatment observations. Conversely, *Silicon Subjects* are the simulated individuals whose monthly states and decisions form the analytic panel. The experiment preserves the same subject population and environment rules across treatment arms while isolating the policy intervention

or shock as the sole source of variance. As illustrated in Figure 1, the workflow proceeds in two stages within an iterative research cycle. Human researchers and Silicon Scientists first formulate the protocol, after which Silicon Subjects reproduce the benchmark experiments. The resulting evaluation is then used to revise assumptions and hypotheses before the calibrated environment is applied to targeted counterfactual experiments. Findings from those experiments initiate the next cycle of evaluation and refinement.

3.1 Silicon Subject Representation

Each LLM agent represents a low-income social individual rather than a single-turn survey respondent. Agent i at month t is represented as

$$\mathbf{s}_{i,t} = [\mathbf{z}_i, \mathbf{x}_{i,t}, \mathbf{m}_{i,t}, \tau_{i,t}, \mathbf{e}_t], \tag{1}$$

where $\mathbf{z}_i$ denotes fixed background attributes, $\mathbf{x}_{i,t}$ represents evolving household and labor states, $\mathbf{m}_{i,t}$ captures individual memory, $\tau_{i,t}$ specifies the active transfer condition, and $\mathbf{e}_t$ records common environmental conditions such as labor-market opportunity or AI shock intensity. This representation prevents a cash-transfer response from devolving into an unconstrained preference statement. The same payment implies different feasible responses for an agent facing children, debt, high rent, health limitations, or limited time.

Fixed attributes describe agents' baseline characteristics, including age, gender, formal education, household composition, initial occupation, risk preference, and pre-treatment financial position. Dynamic states include current employment and occupation, work hours, earnings, expenditure, debt, savings, food security, stress, health and care constraints, job search, skill levels, willingness to undertake training, and AI-tool use. A memory module logs recent policy exposures, historical choices, and cumulative outcomes, ensuring that current behaviors reflect path-dependent adaptation and past hardships rather than merely reacting to the immediate transfer signal.

3.2 Monthly Decision and Transition Process

The experiment advances in discrete monthly rounds. At the beginning of a round, each subject observes an initialized state vector comprised of its current employment status, household liabilities, transfer assignment, wellbeing indicators, AI exposure, and remembered recent events. In response, the LLM outputs a structured decision schema comprising labor actions, a household budget, and wellbeing assessments, accompanied by a brief rationale. The labor component records employment, work hours, job search, training, occupation switching, and AI-tool use; the budget component allocates resources to necessities, care and health, savings, debt repayment, support for others, and discretionary spending.

An accounting layer checks that narrative choices remain economically feasible. Let $y_{i,t}$ be labor earnings, $\tau_{i,t}$ the cash transfer, and $B_{i,t}$ currently available resources. For proposed expenditures $\mathbf{c}_{i,t}$, the system imposes

$$\sum_k c_{i,t}^{(k)} \le y_{i,t} + \tau_{i,t} + B_{i,t} + d_{i,t}^{+}, \tag{2}$$

where $d_{i,t}^{+}$ is allowable additional borrowing. Mandatory expenses, including rent, food, transport, medical needs, and care, are deducted before discretionary spending, and any remaining balance updates savings or debt. This budget constraint prevents agents from reporting mutually incompatible outcomes under limited resources.

Once a decision has been validated, the environment updates the agent:

$$\mathbf{x}_{i,t+1} = F(\mathbf{x}_{i,t}, \mathbf{a}_{i,t}, \tau_{i,t}, \mathbf{e}_t, \epsilon_{i,t}), \tag{3}$$

where $\mathbf{a}_{i,t}$ is the validated action and $\epsilon_{i,t}$ is an idiosyncratic opportunity or hardship component. The labor-market component updates jobs, hours, and earnings; the household-economy component updates consumption, debt, savings, food security, and pressure; and the memory component appends the decision and observed consequence. In the final counterfactual experiment, AI shock modifies F by increasing occupation-specific job-loss and wage-reduction risks, while access to AI-tools expands training and adaptation opportunities but requires agents' time and capacity.

3.3 Policy Modules

Four core modules construct the simulated environment: the transfer module defines cash eligibility and duration; the labor-market module governs wages and occupational transitions; the household-economy module tracks consumption and stress under specific obligations; and the AI-transition module regulates automation risk and retraining access. To evaluate these structural components, we contrast matched simulated populations under an unconditional transfer against a baseline reference regime.

Our AI extension decomposes protection from recovery via two strategic counterfactuals: comparing *UBI + AI Shock* with *AI Shock* to assess cash buffering, while comparing *UBI + AI Shock + AI Tools* with *UBI + AI Shock* to assess the added adaptation support.These contrasts identify latent mechanisms within the synthetic society rather than empirical treatment effects for real-world populations.

3.4 Calibration-First Experimental Logic

The protocol does not interpret novel AI-shock trajectories until the agent environment has been compared with observed UBI regularities. We construct three contextual replicas from parameters specified in the real-study and then summarize the simulated treatment and control trajectories on the outcomes measured in those settings. The calibration stage asks whether simulated agents capture expected directions and pre-specified outcome margins, rather than

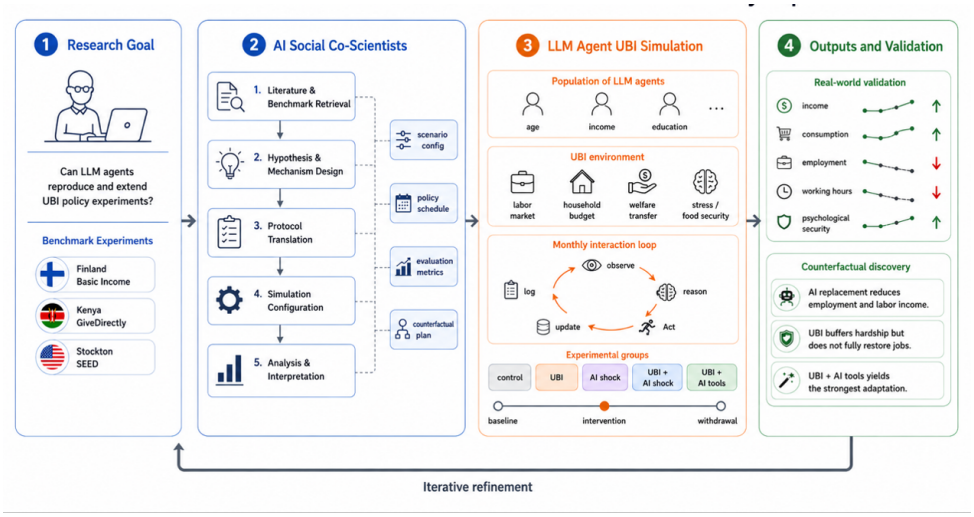

Figure 1. AI Social Co-Scientists and LLM-agent UBI experiment workflow.

claiming perfect numerical reconstruction of the original participants.

Calibration fulfills two roles: it detects simulation anomalies—such as implausible labor supply collapses or artificial health gains—and establishes a validated baseline for counterfactual stress tests across fiscal, household, and environmental dimensions. This framework positions LLM agents as engines for mechanism discovery, predictive calibration, and exploratory policy simulations, while strictly reserving causal claims for empirical field research.

4 Experimental Setting and Design

The three settings differ in welfare institutions, household constraints, and income levels, are not treated as interchangeable examples of "cash." Their different baselines provide a demanding test of whether an agent society can express the same policy through different behavioral pathways.

Table 1 records the core parameters specified in the source study design and used to construct corresponding agent worlds. Correspondence is enforced at the experimental-setting level: eligibility population, transfer amount, duration, macro context, baseline income and consumption, and focal outcomes are mapped into the simulation. The agents remain synthetic abstractions, rather than digital copies of named real participants. This distinction allows for robust external benchmarking while avoiding the false claims of an LLM's capacity to recreate individual human lives.

4.1 Population Construction and Context

The synthetic population is initialized proportionally to match the joint demographic and statistical distributions of each benchmark setting, exhibiting rich heterogeneity across age, gender, education, income, and occupation. Upon initialization, each agent is endowed with a unique personality, familial background, baseline preference, binding constraint matrix, and dynamic memory module. Crucially, these initial states are contextualized: Finnish agents enter an unemployment-benefit regime, Kenyan agents operate in a rural, low-income investment environment, and Stockton agents manage an urban household budget. By powering this architecture with ChatGPT-5.5 and GLM-4.5, the model layer enables cross-backend behavioral comparisons, ensuring results are not artifacts of a single LLM.

For each benchmark, treatment and control populations share identical initialization logic and environmental rules, isolating the cash transfer as the sole source of variance. Each month records the four common outcomes—income, consumption, labor participation, and psychological security—along with mechanism variables such as savings, debt, and training. This high-frequency tracking allows researchers to analyze any given outcome not merely as a static, terminal treatment difference, but as a continuous trajectory shaped by sequential, micro-level household decisions.

4.2 Temporal Protocol and Classic Replicas

Time advances monthly through three conceptually separate phases. The *baseline phase* establishes pre-treatment behavior and verifies that treated and comparison groups begin from comparable trajectories. The *intervention phase* delivers the scenario-specific unconditional transfer to the Sim UBI group while Sim Control retains the reference policy. The *withdrawal/follow-up phase* removes the additional payment and records whether income, consumption, working behavior, and wellbeing return toward control or remain separated through accumulated savings, debt, or adaptation.

The phase design makes two replications possible. A cross-sectional replication compares simulated and real experimental outcomes on summary metrics at meaningful reporting

Table 1. Real-world UBI benchmark settings and experimental parameters.

Scenario	Real-world target population	Intervention and duration	Macro setting	Agent baseline	Scenario-specific focus
Finland	2,000 unemployed benefit recipients, aged 25–58	$560/month, 24 months; control retains conventional benefit	13% unemployment context	Income $1,473/month; consumption $1,076/month	Employment intention and psychological security
Kenya GiveDirectly	Rural low-income, agriculture-oriented, multi-person households; more than 20,000 households	$140/month plus $10 green voucher, 36 months; control receives no transfer	24% unemployment context	Income $182/month; consumption $142/month	Food variety and self-employment
U.S. Stockton SEED	125 urban low-income residents	$500/month plus $25 green voucher, 24 months; control receives no transfer	17% unemployment context	Income $1,450/month; consumption $1,004/month	Consumption upgrading and psychological stress

horizons. A dynamic replication tests whether the simulated society displays credible policy timing: no artificial pre-treatment separation, a response when payments begin, and attenuation or persistence after payments end. The latter requirement is particularly important for LLM agents, which might otherwise repeat a generic belief that cash is beneficial without representing timing or budget dynamics.

After grounding the simulation in the classic settings, we extend the experiment to a prospective AI labor-market disruption. The counterfactual society represents low-income workers in occupations with different AI exposures. A replacement shock begins after an initial observation period and increases employment loss or wage pressure in affected jobs. Table 2 specifies the experimental arms and the mechanism each comparison is intended to reveal.

4.3 Outcomes and Mechanism Tests

Our replication benchmarks cover four outcomes shared across settings: income, consumption, labor participation, and psychological security. These constructs are further operationalized through context-specific measures—employment and distress in Stockton, employment and perceived security in Finland, and assets, self-employment, and food security in Kenya. Agreement with these benchmarks is treated as calibration evidence rather than proof of predictive validity; material deviations require recalibration before counterfactual analysis.

In the AI-shock experiment, trajectories of employment, earnings, stress, and training allow us to distinguish short-run welfare protection from longer-run labor-market adaptation. Income support may ease stress without restoring employment, while training may promote adaptation while anxiety remains elevated. Accordingly, the combined arm pairs income support with AI-tool access and skills training to address both psychological security and adaptation capacity. By varying these components independently, the simulation helps examine candidate mechanisms and generate testable hypotheses for field settings where jointly randomizing them would be costly or infeasible.

5 Experimental Results

5.1 Reproduction of Classic UBI Experiments

Figure 2 compares LLM-agent simulation outcomes with the corresponding real-experiment results. The comparison is deliberately cross-context. A simulation that reproduced only one setting might simply encode the expected answer for that trial, whereas similar directional agreement across Finland, Stockton, and Kenya tests whether the agent environment can retain distinct income levels and institutional settings while responding coherently to unconditional cash.

At the aggregate level, the LLM-agent bars and real-experiment bars are close for average income and consumption in all three scenarios. This matters because the nominal scale of resources differs markedly across the settings: Kenyan household values are much lower than those in Finland and Stockton, yet the simulated society does not mechanically translate the U.S. response into the Kenyan setting. Labor participation remains high and close to the reference results rather than displaying a dramatic simulated withdrawal from work. Psychological security also follows the benchmark direction, including the comparatively weaker Kenyan value relative to the higher security levels shown in Finland and Stockton.

These results should not be interpreted as evidence that simulated agents are behaviorally equivalent to real participants. Instead, they establish a necessary, though not sufficient, calibration criterion for subsequent counterfactual analysis: the architecture reproduces the direction and contextual variation of central cash-transfer outcomes without generating implausible labor-market exits or uniform wellbeing responses. Passing this check supports the use of the model for mechanism exploration.

Figure 3 then reports scenario-specific calibration margins rather than only the shared bar-chart outcomes. The simulation is especially close on U.S. Stockton income and first-year distress (margins 0.19 and 0.19), Finnish second-year employment (0.16), and Kenyan self-employment (0.02). These margins give the replication a gatekeeping role: they indicate where the simulated society resembles the empirical target closely enough to support stress tests and where

additional calibration would be needed before making a substantive interpretation.

5.2 Dynamic Policy Evolution

Static agreement with empirical benchmarks does not establish whether the agents respond coherently to policy timing. We therefore examine trajectories across the baseline, intervention, and withdrawal phases. Credible dynamics require comparable pre-treatment paths, divergence after transfers begin, and either convergence or persistent effects after withdrawal. Such persistence may arise through accumulated savings, debt reduction, or labor-market adjustment. Figure 4 illustrates this test for the Finnish replica.

Before month 13, Sim UBI and Sim Control remain close across the displayed outcomes. When intervention begins, income and consumption in the UBI arm rise immediately. Psychological security also rises markedly, consistent with cash reducing financial uncertainty, while working hours do not display the sharp collapse that an over-simplified "cash removes work incentives" simulation would predict. At withdrawal, outcomes move back toward the control trajectory. Thus, the agents respond to policy timing as well as transfer receipt: the intervention has an onset and an attenuation pattern, not merely a static label attached to an agent profile.

Figure 5 extends the temporal check to all three settings. In each replica, the treatment and control trajectories remain similar before the transfer, diverge after payments begin, and converge or narrow again after withdrawal, although the magnitude and volatility of the responses vary across settings. These phase-responsive patterns show that the simulation links behavioral changes to intervention timing rather than merely fitting endpoint averages. We interpret them as evidence of internal dynamic consistency, complementing rather than replacing the external benchmark calibration reported above.

5.3 Experiment under AI Labor Replacement

We next use the calibrated method for a future setting without an established historical counterpart. Figure 6 compares five conditions that isolate the roles of income security, technological disruption, and adaptation resources. The Control and UBI arms represent trajectories without replacement pressure. The AI Shock arm introduces labor-market displacement without substantial cash protection. The UBI + AI Shock arm tests whether income support alone offsets that pressure. Finally, the UBI + AI Shock + AI Tools arm adds access to productive technology and retraining under the same shock.

The trajectories show a clear simulated disruption mechanism. After the AI shock begins in month 7, the AI Shock arm exhibits steep deterioration in employment and labor earnings and a persistent increase in stress. Training in this arm rises only modestly, suggesting that hardship alone does not make adaptation readily feasible. The simulated workers lose both income and the security required to invest consistently in new skills.

The UBI + AI Shock arm separates psychological protection from labor-market recovery. Relative to AI Shock alone, stress is reduced and hardship is buffered. However, employment and labor earnings remain substantially below the trajectories without displacement and broadly follow the damaged labor-market path. This result is substantively important as a hypothesis: unconditional cash can protect consumption capacity and wellbeing without repairing a demand-side loss of productive work. In this counterfactual, cash acts primarily as a survival guarantee rather than an automatic reskilling policy.

The UBI + AI Shock + AI Tools condition presents a markedly different adaptation trajectory. Employment and income stabilize much nearer to the control path, and training uptake is substantially higher than in the remaining arms. The contrast suggests a complementarity mechanism: cash can provide the short-run room to act, while AI tools and retraining provide an actionable route through which that room becomes labor-market adaptation. The mechanism also clarifies why training should not be interpreted as unambiguously good: the UBI + AI Shock + AI Tools arm can exhibit strong skill investment while stress remains nonzero, because adaptation itself is undertaken in response to a real perceived threat.

5.4 Mechanisms Beyond Existing UBI Evidence

The counterfactual experiments extend existing cash-transfer mechanisms by making institutional, family, labor-market, and environmental constraints intervenable. Table 2 preserves the mechanism hypotheses and simulated contrasts in the study design. The central extension is that cash does not automatically become adaptation capacity: benefit taper rules, care constraints, rent absorption, job supply, informal saving networks, and AI-tool availability govern whether it does so.

Rather than merely measuring mean outcomes, this step isolates the specific structural bottlenecks that determine whether cash transfers translate into work, investment, or wellbeing. By varying U.S. institutional constraints, Finnish labor-market friction, and Kenyan savings networks or environmental shocks, the simulation uncovers hidden policy interactions. These contrasts are uniquely valuable because the simulator can manipulate a single, difficult-to-randomize mechanism while holding the underlying calibrated environment completely fixed.

Taken together, the mechanism contrasts distinguish liquidity from the conditions that make liquidity usable. In Stockton, benefit rules, care burdens, payment timing, housing costs, and access to training shape whether transfers support employment, balance-sheet repair, or consumption smoothing. In Finland, administrative burden primarily affects perceived security, whereas job availability constrains

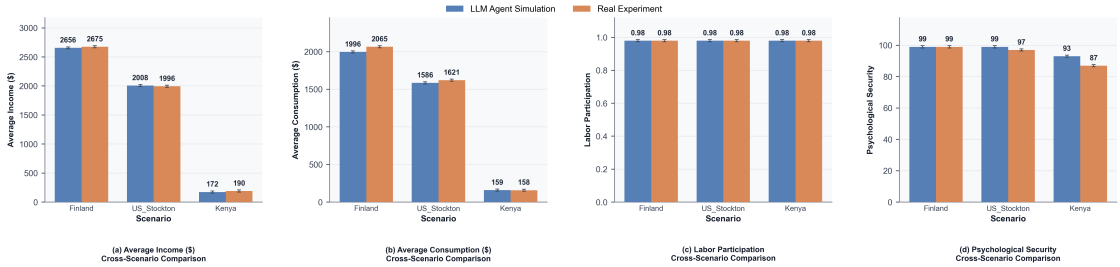

Figure 2. LLM Agent simulation and real-experiment comparison across the three benchmark scenarios.

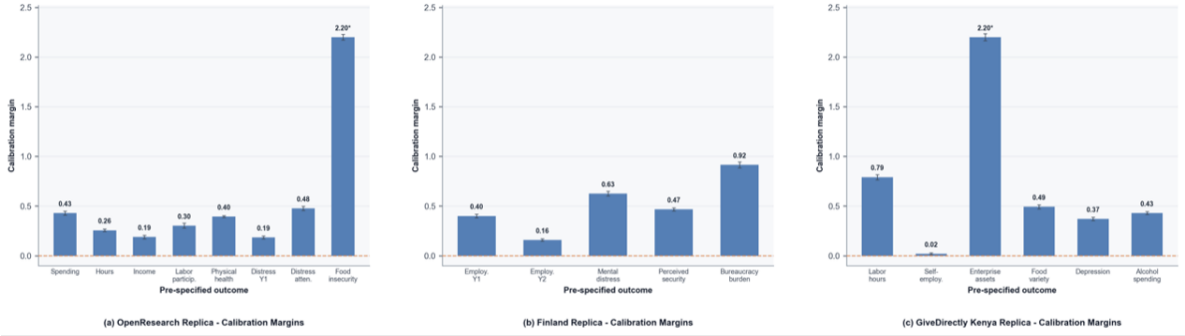

Figure 3. Calibration margins for scenario-specific outcomes in the classic trial replications.

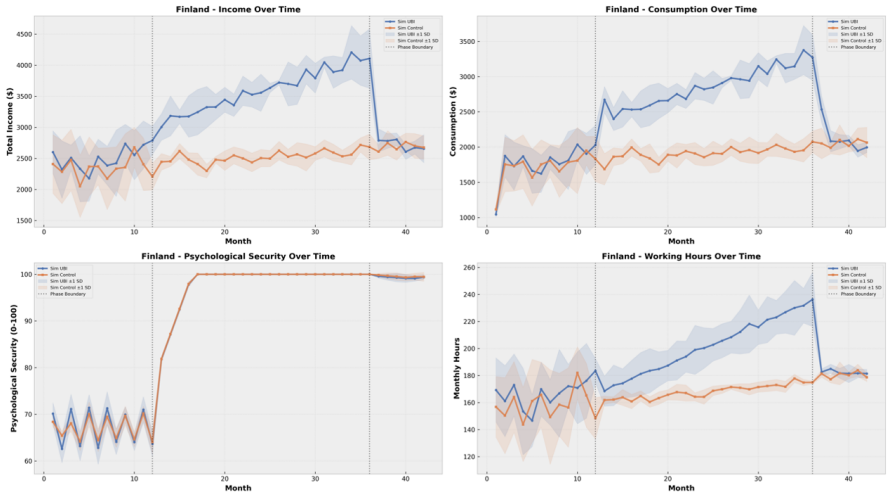

Figure 4. Finnish UBI experiment trajectories over the baseline, intervention, and withdrawal phases.

employment. In Kenya, informal finance, transfer timing, and shared climate or market shocks influence whether cash supports investment, current consumption, or resilience.

Across these settings, cash is neither an work disincentive nor a response to structural disruption. Its effects depend on the capabilities that convert additional resources into feasible action. The AI-shock experiment extends this pattern: UBI primarily buffers hardship, whereas AI tools training under the same shock expand adaptation opportunities. The stronger response of the combined arm should be treated as a mechanism hypothesis for empirical testing rather than as a forecast for a particular population.

6 Conclusion

This study develops a calibration-first workflow for LLM-agent policy experiments in AgentSociety. By separating Silicon Scientists, which assist with research design and interpretation, from Silicon Subjects, which generate the

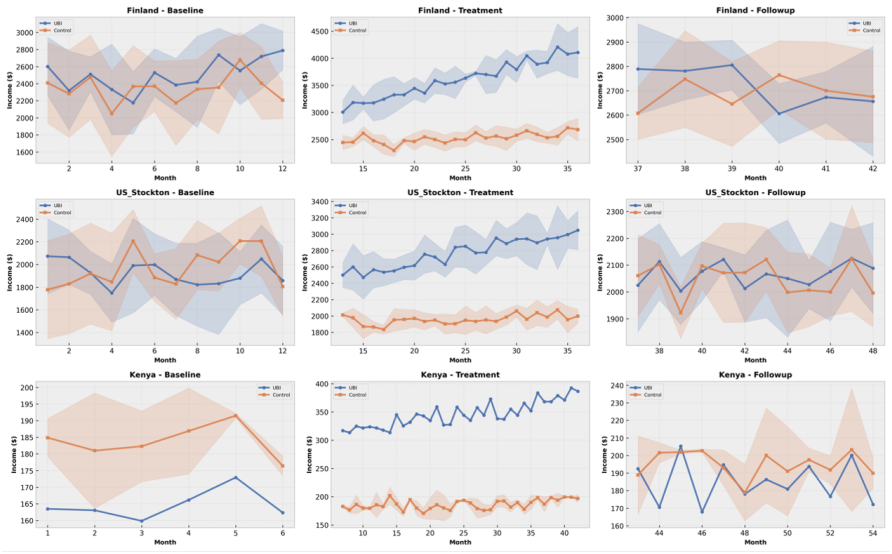

Figure 5. Phase-by-phase income comparison for Finland, Stockton, and Kenya.

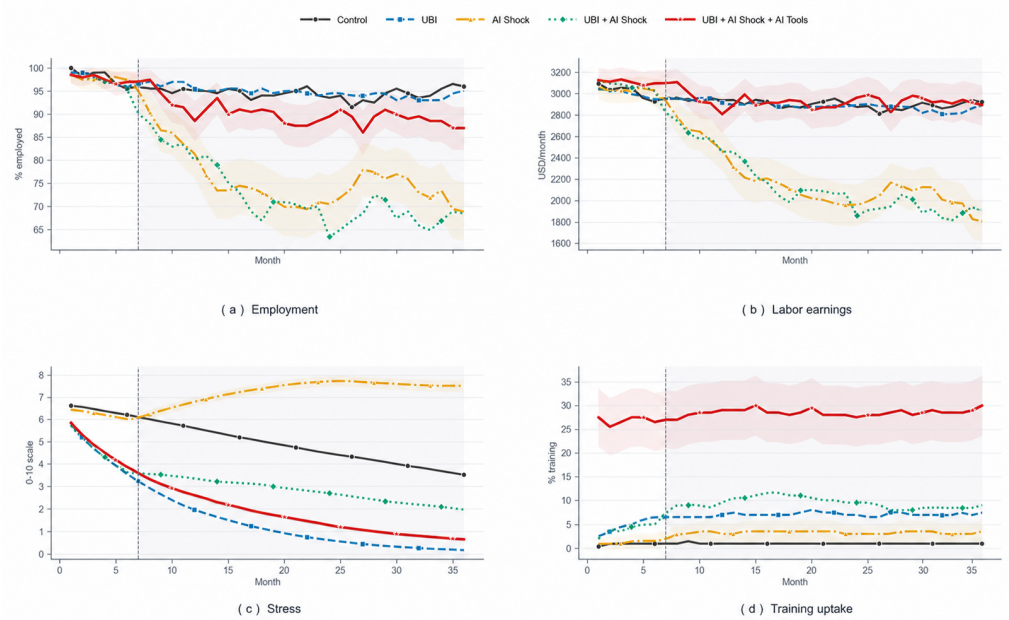

Figure 6. AI labor-market shock trajectories for UBI and AI Shock.

behavioral observations, the workflow makes the roles and assumptions of agent-based experimentation explicit. It first evaluates the simulated environment against cash-transfer evidence from Finland, Stockton, and Kenya before applying the calibrated environment to an AI labor-replacement counterfactual.

Across the three benchmark settings, the simulated outcomes align with key directional and context-specific patterns and exhibit coherent responses to policy timing, satisfying a limited calibration criterion. Under the simulated AI shock, UBI reduces immediate hardship and psychological distress but does not by itself restore employment and earnings. Adding AI-tool access and skills training under the same shock produces a stronger adaptation response.

The contribution is therefore methodological rather than predictive: LLM-agent experiments provide a repeatable environment for probing mechanisms, testing internally consistent counterfactuals, and prioritizing questions for field research. Their usefulness nevertheless depends on benchmark quality, agent design, and robustness across models

Table 2. Prior mechanism evidence and simulated counterfactual contrasts.

Mechanism and setting	Mechanism in prior studies	Simulation result relative to benchmark	Extension generated by this experiment
Remove benefits taper, U.S. Cash	Cash mildly reduces hours; relaxing conditionality alone need not raise employment.	Hours +1.18/week; labor income +$81.79/month; food insecurity -0.77 pp.	Separates cash amount from welfare cliffs: relaxing taper may restore labor more than raising cash.
Cash plus childcare, U.S. Cash	Care support can raise labor supply for caregivers.	Hours +0.44/week; labor income +$12.42/month; training -0.69 pp.	Predicts concentration among high-care-burden households rather than universal training gains.
Annual lump sum, U.S. Cash	Lump sums favor investment; monthly payment smooths consumption and wellbeing.	Debt -$515.97; savings +$3,581.65; food insecurity +1.50 pp; stress +0.14.	Transfers payment-timing insight to indebted U.S. households and reveals a balance-sheet/hardship trade-off.
Rent shock, U.S. Cash	Food, rent, and transport absorb cash expenditure.	Housing +$49.55; stress +0.14; food insecurity +0.46 pp.	Converts housing from a spending channel into a shock that partially absorbs UBI protection.
Cash plus AI tools/training, U.S. Cash	Generative AI may raise productivity, especially for less experienced workers.	Training +19.73 pp; AI use +68.53 pp; labor income +$33.85/month; stress +0.46.	Identifies adaptation pressure: uptake and income can improve without lower stress.
Administrative friction and job supply, Finland	Basic income improves security, while employment gains are limited.	Remove administrative shock: burden -0.95, security +0.15, employment -0.93 pp; raise part-time supply: employment +9.81 pp.	Separates wellbeing gains from vacancy constraints.
Remove ROSCA, Kenya long-term UBI	Saving groups and credit constraints may mediate investment.	Enterprise assets -22.46; savings -13.38; ROSCA participation -38.50 pp.	Makes the informal finance network a switchable mechanism for productive investment.
Hybrid tranching, Kenya short-term UBI	Prior work contrasts monthly and lump-sum transfers.	Enterprise assets +102.25; savings +14.58; food variety -0.54; depression +0.42.	Proposes a hybrid policy that may increase investment while weakening monthly smoothing.
Climate/market risk, Kenya long-term UBI	Local spillovers and prices matter, but common shocks are not randomized.	Climate shock: income -35.30, food insecurity +5.78 pp; saturation: income -2.54, assets -3.98.	Extends UBI evaluation to resilience and local market capacity.

and experimental specifications. Future work should test this robustness systematically and compare simulated predictions with newly observed evidence. LLM-agent policy experiments should thus complement, rather than replace, empirical causal research.

References

[1] Daron Acemoglu and Pascual Restrepo. 2020. Robots and Jobs: Evidence from US Labor Markets. *Journal of Political Economy* 128, 6 (2020), 2188–2244. doi:10.1086/705716

[2] Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, Honolulu, Hawaii, USA, 337–371.

[3] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (2023), 337–351. doi:10.1017/pan.2023.2

[4] David H. Autor. 2015. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives* 29, 3 (2015), 3–30. doi:10.1257/jep.29.3.3

[5] Sarah Baird, Craig McIntosh, and Berk Özler. 2011. Cash or Condition? Evidence from a Cash Transfer Experiment. *The Quarterly Journal of Economics* 126, 4 (2011), 1709–1753. doi:10.1093/qje/qjr032

[6] Abhijit Banerjee, Michael Faye, Alan Krueger, Paul Niehaus, and Tavneet Suri. 2023. *Universal Basic Income: Short-Term Results from a Long-Term Experiment in Kenya.* Technical Report. Abdul Latif Jameel Poverty Action Lab. https://www.povertyactionlab.org/sites/default/files/research-paper/WP4795_Universal-Basic-Income-in-Kenya_Banerjee-et-al_Sept2023.pdf

[7] Alexander W. Bartik, Elizabeth Rhodes, David E. Broockman, Patrick Krause, Sarah Miller, and Eva Vivalt. 2024. *The Impact of Unconditional Cash Transfers on Consumption and Household Balance Sheets: Experimental Evidence from Two U.S. States.* Technical Report 32784. National Bureau of Economic Research. doi:10.3386/w32784

[8] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. 2025. Generative AI at Work. *The Quarterly Journal of Economics* 140, 2 (2025), 889–942. doi:10.1093/qje/qjae044

[9] Dennis Egger, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker. 2022. General Equilibrium Effects of Cash Transfers: Experimental Evidence from Kenya. *Econometrica* 90, 6 (2022), 2603–2643. doi:10.3982/ECTA17945

[10] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2024. Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives. *Humanities and Social Sciences Communications* 11 (2024), 1259. doi:10.1057/s41599-024-03611-3

[11] Johannes Haushofer and Jeremy Shapiro. 2016. The Short-Term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya. *The Quarterly Journal of Economics* 131, 4 (2016), 1973–2042. doi:10.1093/qje/qjw025

[12] John J. Horton. 2023. *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* Technical Report 31122. National Bureau of Economic Research. doi:10.3386/w31122

[13] Damon Jones and Ioana Marinescu. 2022. The Labor Market Impacts of Universal and Permanent Cash Transfers: Evidence from the Alaska Permanent Fund. *American Economic Journal: Economic Policy* 14, 2 (2022), 315–340. doi:10.1257/pol.20190299

[14] Olli Kangas, Signe Jauhiainen, Miska Simanainen, and Minna Ylikännö. 2020. *Evaluation of the Finnish Basic Income Experiment.* Technical Report. Ministry of Social Affairs and Health, Finland. https://tietotarjotin.fi/en/publication/116455/evaluation-of-the-finnish-basic-income-experiment

[15] Sarah Miller, Elizabeth Rhodes, Alexander W. Bartik, David E. Broockman, Patrick Krause, and Eva Vivalt. 2024. *Does Income Affect Health? Evidence from a Randomized Controlled Trial of a Guaranteed Income.* Technical Report 32711. National Bureau of Economic Research. doi:10.3386/w32711

[16] Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* 381, 6654 (2023), 187–192. doi:10.1126/science.adh2586

[17] 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*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3586183.3606763

[18] Jinghua Piao et al. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. arXiv:2502.08691 [cs.MA] https://arxiv.org/abs/2502.08691

[19] Eva Vivalt, Elizabeth Rhodes, Alexander W. Bartik, David E. Broockman, Patrick Krause, and Sarah Miller. 2024. *The Employment Effects of a Guaranteed Income: Experimental Evidence from Two U.S. States*. Technical Report 32719. National Bureau of Economic Research. doi:10.3386/w32719 Revised January 2026.

[20] Stacia West, Amy Castro Baker, Sukhi Samra, and Erin Coltrera. 2021. *Preliminary Analysis: SEED's First Year*. Technical Report. Stockton Economic Empowerment Demonstration. https://www.stocktondemonstration.org/