Grounding the Map of Causality: Why Simulating Behavior Is Not Enough, and What to Measure Instead

Avi Yashchin
Subconscious.ai, New York, NY
avi@subconscious.ai

July 2026 — working paper, not peer reviewed

Abstract

The cost of simulating human behavior with large language models has collapsed. It is now cheap to build a population of synthetic agents and watch them act. This has produced a wave of impressive demonstrations and very little knowledge, because watching an agent behave plausibly tells you nothing about whether it behaves like a person. We reviewed 381 papers that use language models as stand-ins for human subjects. Of those, 104 compare model output to real human data, 24 test the effect of an intervention rather than a static distribution, and four conclude the model recovers the human effect well. One paper in 95 grounds the causal claim it is built on. The missing ingredient is grounding: a comparison to what real humans actually do under the same conditions. Among the studies that do compare, about three quarters check only whether synthetic output matches a static distribution of human responses, and success on the harder interventional test is rare. We argue that the questions worth answering are causal, that a distribution match does not answer a causal question, and that the field should be judged on interventional fidelity, the degree to which a synthetic population reproduces human treatment effects with confidence intervals on held-out experiments. We describe a benchmark for it, connect it to our own results on estimated choice parameters, 0.55 rank correlation across roughly 300 replicated studies and 0.73 on the 43 that pass design filters, and lay out a fidelity hierarchy from sign to calibrated effect size. As a standing commitment, we will report interventional fidelity, CI-overlap on held-out human experiments, as we run them, including the misses. The goal is not a better demo. The goal is a map of cause and effect you can actually navigate by.

Keywords: causal inference, synthetic respondents, discrete choice experiments, large language models, treatment effects, validation, agent-based simulation

1 Simulation is cheap. Knowing is not.

As the cost of building anything drops toward zero, the value of knowing what to build rises toward infinity. Large language models have driven the cost of simulating a person close to zero. You can spin up a thousand synthetic respondents, give them personas, and watch them answer surveys, trade in markets, or form a society (Park et al., 2023). The demonstrations

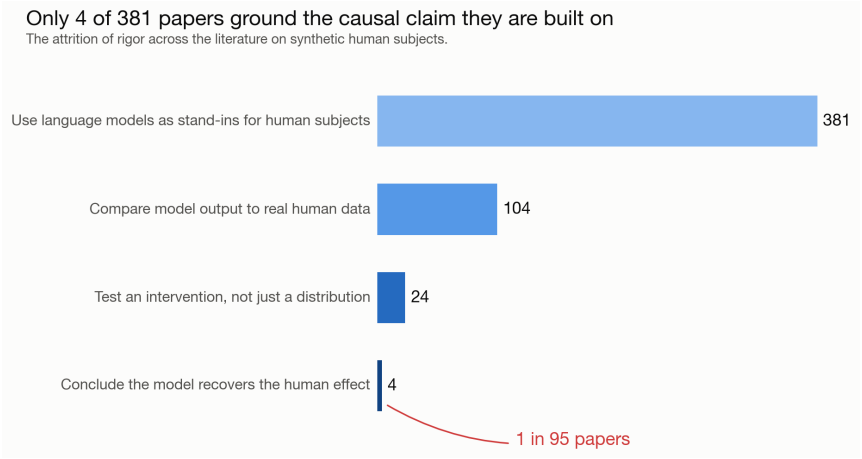

Figure 1: The attrition of rigor. Of 381 papers that use language models as stand-ins for human subjects, 104 compare their output to real human data, 24 test the effect of an intervention rather than a static distribution, and 4 conclude the model recovers the human effect well. One paper in 95 grounds the causal claim it is built on.

are striking. Agents gossip, throw parties, run for office. Long-horizon agent laboratories now score autonomy over hundreds of simulated tasks.

Almost none of it is grounded. Watching an agent do something plausible is not evidence that a human would do the same thing. A society of agents that behaves like a society is a mirror, not a measurement, until you check it against people. The generative-agent studies that started this wave are explicit that their contribution is believable behavior, not correspondence to human data (Park et al., 2023). That is an honest framing and a real engineering achievement. It is also the exact place where the field stops and the useful work should begin.

The useful work is causal. A company does not want to know that a synthetic shopper acts like a shopper. It wants to know whether a two-dollar price increase will cost it sales, whether one message moves behavior more than another, whether a policy will land the way it was designed to. These are questions about what happens when you change something. Judea Pearl calls this the difference between seeing and doing (Pearl and Mackenzie, 2018; Pearl, 2009). Watching agents behave lives on the bottom rung of his ladder, association. The decisions that matter live on the second rung, intervention. A simulation that never climbs the ladder cannot answer the questions people are paying it to answer.

This paper makes three claims. The field validates synthetic humans overwhelmingly by association, not intervention (Section 2). Association is not enough, because a model can match a distribution and still get the effect of a change wrong (Section 3). And there is a concrete standard that fixes this, interventional fidelity, along with a benchmark to measure it and a path we have already partly walked (Section 4).

2 What the field actually measures

We assembled a corpus of 381 papers on language models as stand-ins for human subjects and found 104 that report a real comparison of model output against human data. We coded

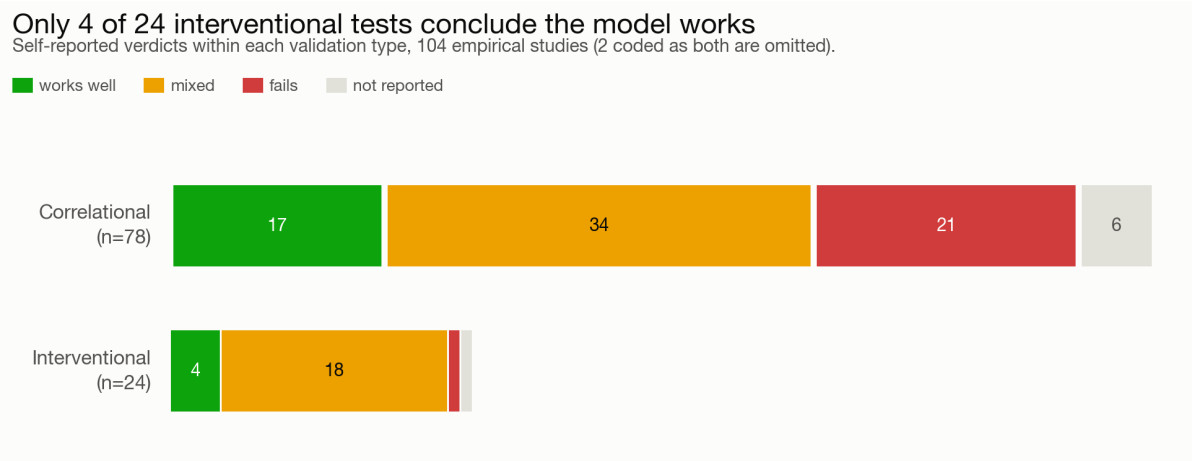

Figure 2: Verdict within each validation type across 104 empirical studies. Interventional validation reaches "works well" in four studies of twenty-four. Validation type was coded from each paper's title and abstract against a fixed rubric; the verdict is the paper's own self-reported conclusion. Both codes, with a one-line rationale per study, ship with the evidence-table release described in the acknowledgments.

each by how it validates the model. A study is correlational if its main comparison is between synthetic output and a static human distribution: matching opinion-poll marginals, survey response distributions, personality scores, or labeled-data accuracy. A study is interventional if it compares the effect of a manipulation, a treatment, a set of varied attribute levels, a price or framing condition, between the model and real people. In plain terms: a correlational study asks whether the synthetic crowd *looks like* people, and an interventional study asks whether it *responds to a change* like people. Only the second answers the question a decision-maker is actually paying to answer.

The corpus and the coding are deliberately simple, and their limits belong in the text. The corpus comes from a keyword sweep of arXiv, cross-checked against two curated reading lists from our research community; journal-only and SSRN work is catalogued separately and not yet coded, so fields that publish off arXiv are underrepresented. The sweep errs toward inclusion, so the 381 denominator carries adjacent work at its edges. Validation type was coded from each paper's title and abstract against a fixed rubric by a language model, and the verdict is each paper's own self-reported conclusion. Neither choice props up our argument: a looser denominator inflates only the top of the funnel, while the ratios the argument rests on, the interventional share of the 104 grounded studies and the verdicts within it, do not touch the denominator at all, and self-reported verdicts flatter the field rather than us.

The result is lopsided. Roughly 75% of the studies are correlational and about 23% are interventional, with the rest doing both. Figure 2 shows the verdict within each type, and it is the part that should give the field pause. Among the interventional studies the most common verdict is "mixed," and only four of twenty-four conclude the model works well. Reproducing a human effect is harder than reproducing a human distribution, and the evidence that anyone can do it reliably is thin.

The weakness is not confined to one field. Figure 3 breaks the empirical studies down by domain. No domain reaches a "works well" verdict in more than a third of its studies, and

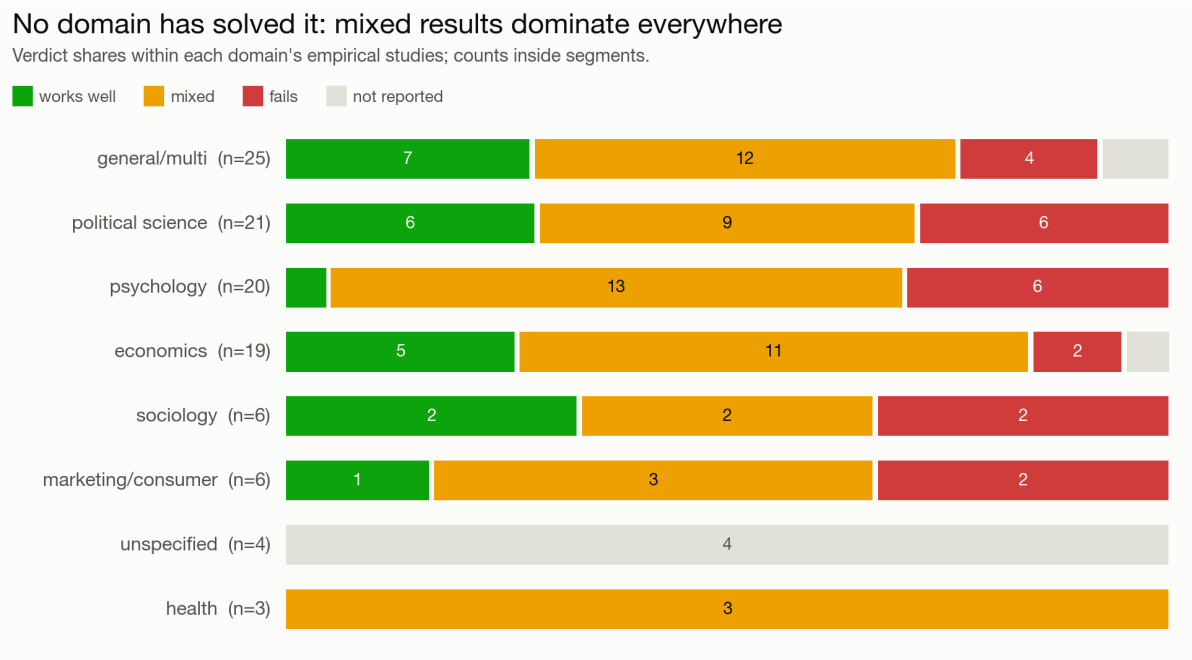

Figure 3: Verdicts by domain, as a share of each domain's empirical studies. Green is "works well," yellow "mixed," red "fails." Domain sizes (n) are small in several fields and the shares should be read with that in mind, but the pattern is consistent: the modal outcome everywhere is a mixed result.

psychology ($n = 20$), a field where synthetic respondents are especially tempting to deploy, reaches it in only one study of twenty.

The gap is not closing on its own. Figure 4 tracks the interventional share of new empirical work by year: it sat near a quarter in 2024 and 2025 and, in the partial 2026 data, none of fifteen studies is interventional. The field is adding volume at the bottom of the ladder faster than it is climbing.

The founding papers of this area are, correctly, correlational. Argyle and colleagues introduced "algorithmic fidelity," conditioning a model on demographics and measuring how well its response distributions track human subgroups (Argyle et al., 2022). Santurkar and colleagues asked whose opinion distributions a model reflects (Santurkar et al., 2023). This work established that a model can resemble a population. It did not establish that a model will respond to a change the way that population would, and it did not claim to.

The interventional minority is where the field is really tested, and the results are uneven. Studies that put models into manipulated experiments, trust games (Xie et al., 2024), ultimatum and strategic settings (Sreedhar and Chilton, 2024), market experiments with feedback (del Rio-Chanona et al., 2025), classic behavioral economics (Horton, 2023), and classic cognitive manipulations (Binz and Schulz, 2022; Aher et al., 2022), tend to find that models get the direction of an effect right and the size, the spread, or the split across subgroups wrong. The largest replication efforts land the same way: across 156 scenario-based experiments (Cui et al., 2024) and 133 media-effects findings (Yeykelis et al., 2024), models recover the direction of roughly three quarters of main effects and are markedly weaker on interactions and

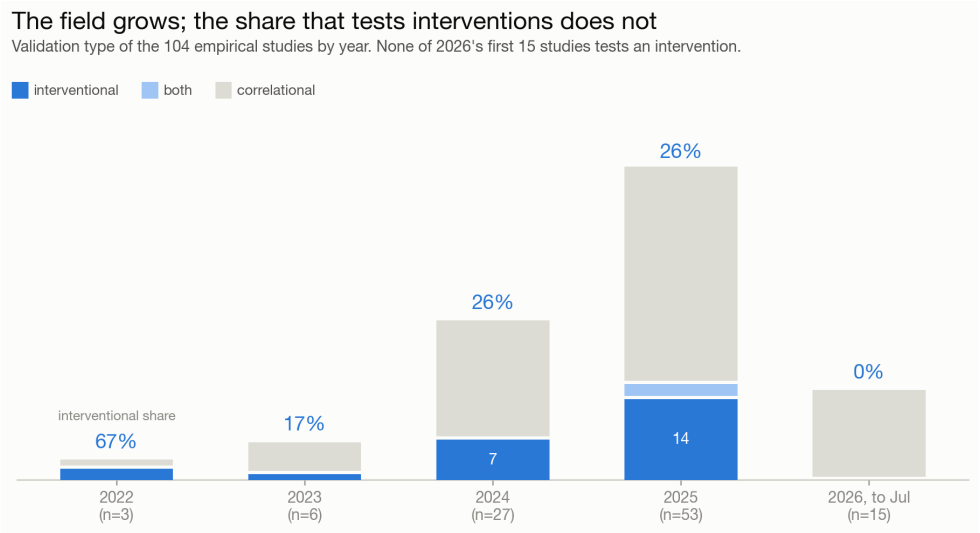

Figure 4: Interventional share of empirical LLM-versus-human studies by year. The absolute count of interventional work is not rising as a fraction of the field, and 2026 (partial) so far contains none.

magnitudes.

3 Why matching a distribution is not enough

Every applied use of a synthetic respondent is a causal question in disguise. Choosing between two prices, two messages, or two product configurations is asking what behavior does when the chosen variable changes. That is a treatment effect, a contrast between two conditions, not a feature of one distribution.

A model can nail the distribution and miss the effect. Suppose synthetic and human respondents agree on how much, on average, they like a product. It does not follow that they agree on how much a two-dollar price increase cuts purchase intent. The first quantity is a marginal. The second is a difference between two marginals under two conditions, and errors that cancel in the average need not cancel in the difference. This is why correlational fidelity is necessary but not sufficient for causal use, a point the field's own critics make when they show a model reproducing surface responses while diverging on the mechanism underneath (Wang et al., 2024; Gao et al., 2024).

The known failure modes all bite harder on effects than on distributions.

- **Variance collapse.** Synthetic respondents pile onto the reasonable middle (Petrov et al., 2024), erasing the heterogeneity that decides whether an effect is broad or concentrated in one segment.
- **Demographic flattening.** Identity groups get homogenized or misportrayed (Wang et al., 2024), so an effect can be right on average and wrong for every group a decision-maker cares about.

- **Over-rationality.** Models drift toward textbook-optimal play and mute the anomalies, loss aversion, framing, anchoring, that interventions are designed to exploit.
- **Prompt sensitivity.** Results move with wording, so an unstated design choice can manufacture or erase an effect.
- **Social-desirability bias** (Salecha et al., 2024) that shifts levels, and can shift differences, along the same axis as the treatment.

There is a competitive point buried here, and it is bigger than one startup's positioning. The market-research and experience-management industry, survey platforms, panel providers, brand trackers, sells correlational evidence: who says what, how satisfied people report being, how a distribution of opinions moved this quarter. Those dashboards are useful, and every one of them lives on the bottom rung of the ladder. "Our synthetic users answered a survey" and "our simulation correlates with history" are claims anyone can match by lunchtime, and neither one tells you what happens when you change the price. The claim that is hard to make, and worth making, is narrow: this synthetic population recovers the causal effect of the decision in front of you, with a stated confidence interval, checked against a randomized human experiment. The incumbent industry cannot currently be graded on that standard at all, because its methods do not produce counterfactuals. We propose to be graded on it in public, misses included.

4 Interventional fidelity, and how to measure it

We propose that the field be judged on interventional fidelity: whether a synthetic population reproduces human treatment effects, reported with confidence intervals, on held-out randomized experiments. This is a strictly harder target than distributional match, because an effect is a contrast between conditions and getting a contrast right demands getting both conditions right in the way they differ. It also puts the metric where the decision is.

This is an extension of a method we have run, not a proposal on paper. In Yashchin et al. (2024) we built a discrete-choice pipeline, the survey method, usually called conjoint analysis, behind most pricing and product research: a model generates attributes and levels, an orthogonal or full-factorial design fixes the experiment, synthetic personas are drawn to match population marginals from ACS and NHTS, and a McFadden choice model estimates the parameters (Train, 2009). We replicated roughly 300 published stated-preference studies with it. Scored by Spearman rank correlation on estimated choice parameters, not raw responses, we saw 0.55 across all replications and 0.73 on 43 studies after removing order-bias and low-information designs, with per-domain means from 0.54 to 0.98. Rank correlation of estimated parameters is already halfway up the ladder. It asks whether the model orders attribute effects the way humans do. Interventional fidelity finishes the climb, from ordering to magnitude, and from one method to any manipulation.

The Interventional Fidelity Benchmark. The benchmark is a set of published human randomized experiments chosen so that each one (i) reports a treatment effect with a confidence interval or standard error, (ii) manipulates a factor you can express as an attribute, a message, or a price, and (iii) covers a target domain, consumer choice, pricing, public policy, health. To

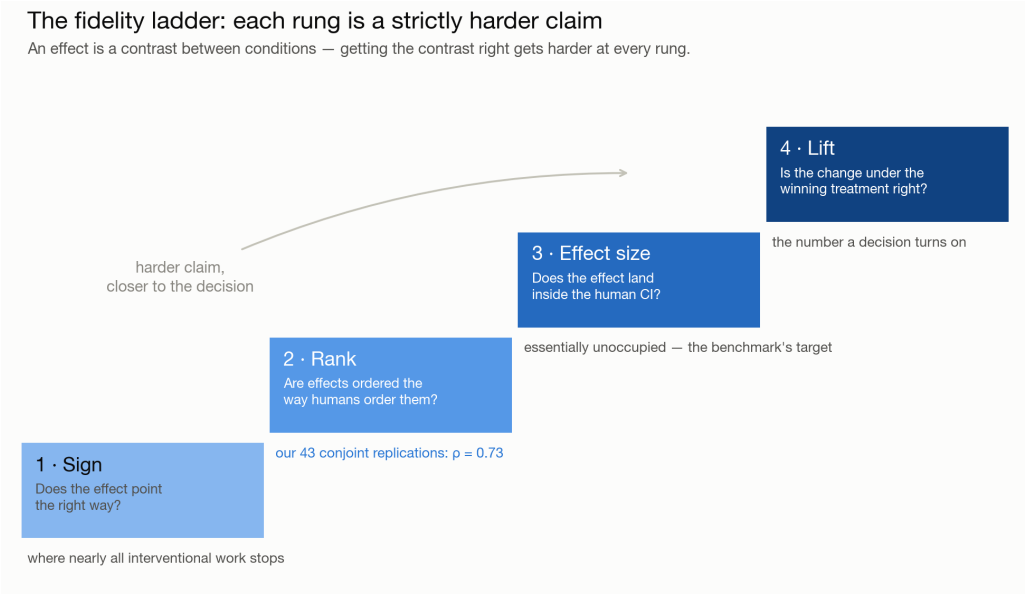

Figure 5: The fidelity ladder. Each rung is a strictly harder claim than the one below it, because an effect is a contrast between conditions. Nearly all interventional work reports only the sign; our conjoint replications reach rank fidelity at 0.73. On the effect-size rung the best prior evidence is pooled, a 0.85 correlation across an archive of attitude experiments (Ashokkumar et al., 2026); per-experiment calibration on decision-grade interventions remains open, and the lift rung is empty.

keep the model from reciting what it memorized, the experiments are drawn from preregistered or held-out sources and, where possible, come from after the model's training cutoff. Each one is re-run as a synthetic experiment through the pipeline above, blind to the human result, and returns a treatment effect with its own confidence interval.

A fidelity hierarchy. We report agreement at four rising levels of difficulty, so the number is legible both to a field that mostly reports direction and to a decision-maker who needs a magnitude (Figure 5).

1. **Sign.** Does the synthetic effect point the right way? This is where most work sits today.
2. **Rank.** Spearman correlation of attribute effects, the same metric that gives us 0.73.
3. **Effect size.** Does the synthetic effect land inside the human confidence interval? Reported as a CI-overlap rate and a calibration plot of synthetic against human effects, each with its interval.
4. **Interventional lift.** Does the model predict the change in outcome under the winning treatment, the exact quantity a decision turns on?

Level three is the line almost no one has crossed, and the strongest result near it sharpens the target rather than closing it. Ashokkumar et al. (2026) show that GPT-4-simulated respondents predict the treatment effects of 70 preregistered survey experiments at a pooled

correlation of 0.85, rising to 0.90 on unpublished studies that cannot sit in the training data. That is real evidence that language models carry causal signal, and any honest map of this field puts it at the top. It is also a different claim from the one a decision needs. A pooled correlation across an archive of brief attitudinal experiments does not say whether *this* synthetic effect lands inside *this* human confidence interval, on the choice and pricing interventions companies act on, before the human result exists. That per-experiment claim, calibrated effect by calibrated effect, misses included, is the one we could not find in the 104 studies or in any competitor's public material, and it is the rung this benchmark grades.

Figure 6 shows what a single point on this ladder looks like in practice, from one replication of the immigrant conjoint of Hainmueller et al. (2014): which attributes of a hypothetical immigrant, education, profession, language, country of origin, make Americans more or less likely to admit them. Across the 22 attribute effects that match between the two designs, the synthetic population orders effects the way humans do at a rank correlation of 0.64, and the misses are systematic rather than random. Education, language, and employment effects track the human estimates closely. The failures concentrate where the model's values diverge from those of 2010-era survey respondents: it inverts the human penalty on unauthorized prior entry and scrambles the country-of-origin preferences. A benchmark that reports this plot, not a single scalar, makes both the fidelity and its failure mode legible, and the failure mode is exactly what a user needs to see before trusting the tool on an adjacent question.

The identification objection. There is a sharper objection to synthetic experiments than any behavioral failure mode. Gui and Toubia (2023) show that varying a treatment inside a prompt shifts the background variables the model silently infers, so a naive synthetic A/B test can violate the unconfoundedness that randomization buys in a human experiment: tell a synthetic shopper the price is higher and it may quietly imagine a different store. Any benchmark in this space has to answer that, not footnote it. Our pipeline answers part of it by construction, because a factorial design varies all attributes jointly under an explicit experimental design and effects are estimated from designed contrasts rather than from one free-text manipulation at a time. The rest is an empirical question the benchmark should measure rather than argue, so it joins the stress tests below.

Stress tests. Each failure mode becomes a test the benchmark reports rather than a caveat it hopes to outrun. Distributional fidelity by Wasserstein and KS distance, not just means, against variance collapse. Per-subgroup effect fidelity against demographic flattening. Behavioral-anomaly experiments, framing and loss aversion, against over-rationality. Effect stability across seeds and prompt variants against prompt sensitivity, which our prior work already handles through an order-invariance filter. And effect agreement between design-blind and design-aware prompting, against the confounding channel of Gui and Toubia (2023).

5 The map is the point

Step back from the metric. A benchmark that certifies synthetic populations recover human treatment effects is not just a validation checklist. It is the unit of construction for something larger: a map of cause and effect across the decisions that markets never price well. Most of what shapes human life, a policy, a message, a design, a norm, has no ticker and no market

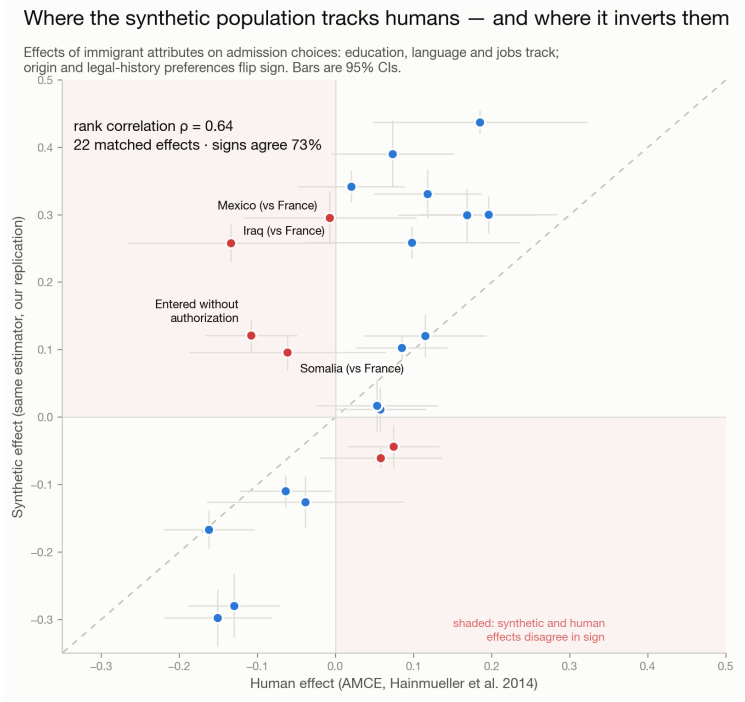

Figure 6: Synthetic against human average marginal component effects (AMCEs) for one replication of the immigrant conjoint of Hainmueller et al. (2014), on the 22 attribute-level effects shared by the two designs. Both sides use the same estimator, a linear probability model on attribute dummies, with 95% confidence intervals; country-of-origin effects are re-referenced to France on both sides because the designs use different baseline countries. Points on the dashed diagonal are perfect recovery; Spearman rank correlation is 0.64 and signs agree on 73% of effects. The sign flips (red) concentrate in origin and legal history. Choice data, matching, and estimation code ship with the replication release described in the acknowledgments. This is a single point on the effect-size rung, and the kind of figure the benchmark produces per study.

to optimize it. The way you learn cause and effect for those things is experimentation, and experimentation on real people is slow, expensive, and often unethical to run at the scale the questions deserve. A grounded synthetic population, one that has earned trust by reproducing human effects where we can check it, lets you run the experiment first in silicon and reserve human studies for confirmation.

That is the difference between a flight simulator and a video game. Both show a plane. Only one is trusted to train a pilot, because only one is validated against physics. Interventional fidelity is how a simulation of people earns the same trust. It is also, we think, the honest version of a much-abused phrase. Causal science that can measure the effect of a change, ethically and at scale, on the parts of life that markets ignore, is worth building carefully. It is not worth faking with ungrounded agents that behave well and mean nothing.

6 Limitations

We are candid about the difficulty, because the data is. Our own coding shows positive interventional results are scarce precisely because effects are hard to reproduce. Effect-size fidelity at level three is a harder target than the rank fidelity where we report 0.73 in aggregate, and single studies land lower: the replication in Figure 6 reaches 0.64, with sign flips we display rather than hide. Early CI-overlap numbers may be lower still. That is acceptable. A calibrated result is credible at moderate overlap, and honest calibration is the foundation of trust that separates a decision tool from a convincing picture. The binding constraint is not modeling. It is curation. Assembling human experiments with usable confidence intervals and manipulable factors is the real labor, and it is exactly what a maintained map of the field contributes to the benchmark.

Acknowledgments

We thank the Subconscious.ai team. The evidence base for this paper, a 381-paper corpus and a 104-study evidence table with per-paper validation-type coding and written rationales, together with the raw synthetic choice data, level matching, and estimation code behind Figure 6 and the scripts that generate the figures, is being prepared for public release alongside a full re-run of the replication suite and will be published with this paper.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. *arXiv preprint arXiv:2208.10264*, 2022. URL `https://arxiv.org/abs/2208.10264`.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*, 2022. URL `https://arxiv.org/abs/2209.06899`.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, 2026. In press. Circulated in 2024 as the working paper "Predicting Results of Social Science Experiments Using Large Language Models".

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *arXiv preprint arXiv:2206.14576*, 2022. URL `https://arxiv.org/abs/2206.14576`.

Ziyan Cui, Ning Li, and Huaikang Zhou. Can large language models replace human subjects? a large-scale replication of scenario-based experiments in psychology and management. *arXiv preprint arXiv:2409.00128*, 2024. URL `https://arxiv.org/abs/2409.00128`. Published version in *Nature Computational Science*, 2025.

R. Maria del Rio-Chanona, Marco Pangallo, and Cars Hommes. Can generative AI agents behave like humans? evidence from laboratory market experiments. *arXiv preprint arXiv:2505.07457*, 2025. URL `https://arxiv.org/abs/2505.07457`.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using LLMs as human surrogates: Scylla ex machina. *arXiv preprint arXiv:2410.19599*, 2024. URL https://arxiv.org/abs/2410.19599.

George Gui and Olivier Toubia. The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*, 2023. URL https://arxiv.org/abs/2312.15524.

Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1):1–30, 2014.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*, 2023. URL https://arxiv.org/abs/2301.07543. A 2026 revision (v2) adds Apostolos Filippas and Benjamin S. Manning as coauthors.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. URL https://arxiv.org/abs/2304.03442.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

Nikolay B. Petrov, Gregory Serapio-García, and Jason Rentfrow. Limited ability of LLMs to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*, 2024. URL https://arxiv.org/abs/2405.07248.

Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*, 2024. URL https://arxiv.org/abs/2405.06058.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023. URL https://arxiv.org/abs/2303.17548.

Karthik Sreedhar and Lydia Chilton. Simulating human strategic behavior: Comparing single and multi-agent LLMs. *arXiv preprint arXiv:2402.08189*, 2024. URL https://arxiv.org/abs/2402.08189.

Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition, 2009.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *arXiv preprint arXiv:2402.01908*, 2024. URL https://arxiv.org/abs/2402.01908.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, et al. Can large language model agents simulate human trust behavior? *arXiv preprint arXiv:2402.04559*, 2024. URL `https://arxiv.org/abs/2402.04559`.

Avi Yashchin, Chi Wing Ng, Harshita Khandelwal, Jennifer H Lee, Mahdi Jafari, et al. Evaluation of large language models for preference elicitation. Technical report, Subconscious.ai, 2024. Internal technical report; full replication release in preparation.

Leo Yeykelis, Kaavya Pichai, James J. Cummings, and Byron Reeves. Using large language models to create ai personas for replication and prediction of media effects: An empirical test of 133 published experimental research findings. *arXiv preprint arXiv:2408.16073*, 2024. URL `https://arxiv.org/abs/2408.16073`.