The Causal Atlas: A Program to Replicate the Behavioral Sciences

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

August 2026 — working paper, not peer reviewed

Abstract

Economics, sociology, and psychology contain a vast record of claims about what changes human behavior. That record is stored as prose, tables, appendices, code, and scattered replication packages. It is read paper by paper and replicated study by study. We propose the *Causal Atlas*: an open, versioned system that turns every eligible intervention study into an executable evidence record, reruns it against synthetic populations, anchors those runs to human results, and publishes the agreement and disagreement. The goal is not to declare the literature replicated by a model. Synthetic replication is a scalable screening and measurement layer; human replication remains the ground truth. Each Atlas entry binds an intervention, population, outcome, estimand, protocol, effect estimate, uncertainty interval, provenance trail, and fidelity history. The current repository contains 154 runnable conjoint-study transcriptions. They are a seed, not coverage of the behavioral sciences and not evidence of corpus-wide fidelity. This paper defines the unit of replication, the six-stage production loop, the validation standard, the inclusion and governance rules, and a path from that seed to a living map of causal evidence. The deliverable is simple: for any claim that changing X changes Y for population P, a reader should be able to inspect the original evidence, execute the study, compare human and synthetic effects, and see what is known, what failed, and what remains untested.

Keywords: replication, causal inference, synthetic populations, behavioral science, experimental economics, computational social science, open science

1 The missing object is a map

The behavioral sciences do not lack studies. They lack a common, executable representation of what those studies claim. A price experiment in economics, a framing intervention in psychology, and a social-norm treatment in sociology may share the same causal form: change a treatment, observe an outcome, and estimate the difference for a population. Yet their protocols and results live in incompatible documents and data formats. A researcher who wants to know whether an effect travels must reconstruct the study before testing it.

Large replication projects have shown both the value and the cost of doing that work carefully. Coordinated efforts in psychology, experimental economics, and the broader social sciences found that replication effects were often smaller than the originals and that replication success varied substantially across studies (Open Science Collaboration, 2015; Camerer et al., 2016, 2018).

Their contribution was not a single verdict on a discipline. It was an auditable comparison between an original effect and a new estimate. The limitation is throughput: each project required teams to rediscover materials, translate protocols, resolve ambiguities, recruit people, and harmonize results.

The Causal Atlas is a proposal to make that work cumulative. It treats a study not as a PDF to be summarized but as a versioned causal object that can be executed, tested, corrected, and connected to related objects. The destination is a living map of claims of the form

$$\text{intervention } X \longrightarrow \text{outcome } Y \quad \text{for population } P,$$

with an effect estimate, an uncertainty interval, and a visible record of how the claim behaves under replication. The map does not erase context. Population, setting, timing, measurement, and experimental role are part of the edge, not metadata that can be dropped.

This is a vision paper. It defines the target and the contract. It does not report a completed corpus-wide replication, and it does not infer that an effect is true because a language model reproduces it.

2 What "every study" means

Every study is a direction of travel, not a claim of present coverage. The operational target is every *eligible causal study* in economics, sociology, and psychology. A study is eligible when it has:

1. a defined intervention or as-if intervention and comparison condition;
2. a defined outcome and estimand;
3. a population and setting that can be represented without inventing identity or context;
4. enough protocol detail to reconstruct treatment assignment and measurement;
5. a recoverable human result against which a rerun can be compared; and
6. an ethical and legal path to storing the specification and executing the task.

Randomized experiments, conjoint and discrete-choice experiments, field experiments, and some quasi-experimental designs can satisfy this contract. Purely descriptive or correlational studies do not enter the causal map until their estimand is made explicit. Studies that depend on unsafe interventions, protected data that cannot be represented, or irrecoverable protocols remain catalogued as unavailable rather than silently approximated.

This boundary matters. Causality concerns interventions, not resemblance (Pearl, 2009). A synthetic population can match a survey distribution and still predict the wrong response to a change. Likewise, a replication can preserve the treatment labels while changing the decision maker, outcome, or measurement context. The Atlas calls those different studies.

3 The atomic unit: an executable evidence record

The unit of the Atlas is neither the paper nor the model response. It is an executable evidence record. Table 1 gives its minimum contract.

Field	Required content
Identity	Persistent study and version identifiers; citation; license; source files and checksums.
Population	Who or what is studied, sampling frame, geography, period, inclusion rules, and relevant covariates.
Design	Assignment mechanism, arms or attribute levels, task flow, randomization, sample size, and stopping rules.
Estimand	Treatment, outcome, unit, contrast, estimator, reference level, and uncertainty procedure.
Execution	Machine-readable protocol, software and model versions, prompts where applicable, seeds, and run logs.
Evidence	Original human estimate, replication estimates, confidence intervals, exclusions, quality flags, and failures.
Provenance	A trace from each encoded value to the paper, supplement, data, code, or an explicit curator decision.

Table 1: Minimum contract for an executable evidence record. Missing fields are reported as missing; they are not filled with plausible values.

This representation changes the economics of replication. Once the protocol and estimand are explicit, the same record can support a computational rerun, a human replication, a different population, a different model, or a reanalysis. Corrections produce a new version instead of rewriting history. The record follows the spirit of FAIR research objects—findable, accessible, interoperable, and reusable—while adding the execution and causal semantics required for experiments (Wilkinson et al., 2016).

For conjoint studies, the estimand may be an average marginal component effect or another choice-model parameter (Hainmueller et al., 2014). For a randomized trial it may be an average treatment effect; for a quasi-experiment, a local effect under explicit identifying assumptions. The Atlas does not force these into one number. It standardizes the questions needed to compare like with like.

4 The production loop

The Atlas is built by repeating one six-stage loop.

1. **Discover.** Search journals, repositories, registries, citations, and replication packages. Deduplicate versions and identify the experiment-level unit.

2. **Formalize.** Extract the population, intervention, outcome, design, estimand, and human result. Link every field to its source. Automated extraction proposes; validation rules and curators decide.

3. **Compile.** Convert the protocol into an executable task. Static checks reject missing arms, impossible assignments, mismatched reference levels, ambiguous units, and identity errors before a participant or model is run.

4. **Execute.** Run the task first on one or more synthetic populations as a cheap, repeatable probe. When authorized and useful, run a preregistered human replication from the same specification.

5. **Score.** Estimate the same causal contrast as the original. Compare sign, rank, magnitude, uncertainty, and heterogeneity. Record design deviations and failed runs rather than removing them from the denominator.

6. **Publish.** Release the record, artifacts, results, and version history. Feed errors back into the specification and rerun without erasing prior evidence.

Natural-language experiment representations have already shown that heterogeneous tasks can be gathered at meaningful scale. Psych-101 encoded 160 experiments and more than ten million trial-level choices for training and evaluating a broad cognitive model (Binz et al., 2025). Other projects have used language models to simulate samples or rerun collections of published experiments (Argyle et al., 2023; Yeykelis et al., 2024). The Atlas adopts the scaling lesson but adds a stricter object: each run must point back to a human causal estimate and a reproducible protocol.

Automation is useful precisely where it can fail visibly. A model can locate candidate papers, transcribe tables, propose mappings, generate executable forms, and flag inconsistencies. It cannot be allowed to invent a baseline, silently substitute a population, or turn an association into an intervention. Every automated transformation therefore emits evidence for review and can be refused.

5 Synthetic replication is a screen, not ground truth

The fastest version of the Atlas reruns studies with synthetic respondents. That is not the same as replicating the study with people. It is a computational replication of the protocol and a test of a particular simulation system. Its value is scale: it can reveal broken transcriptions, identify effects a model gets systematically wrong, compare model versions, and prioritize expensive human work.

The epistemic order must remain clear:

1. the original human study supplies a claim and an initial estimate;

2. synthetic reruns measure whether a simulator recovers that human effect;

3. independent human replications measure whether the effect itself travels; and

4. the Atlas accumulates all three without treating any one as infallible.

Synthetic fidelity is evaluated on a ladder. The lowest rung asks whether the effect has the same sign. The next asks whether multiple effects have the same ordering. Higher rungs ask whether magnitudes are calibrated, uncertainty intervals overlap under comparable designs, and a held-out intervention produces the same lift. No single correlation is a certificate of validity. Scores are conditional on study, population, model, prompt, design, and time.

This separation creates a useful feedback system. Studies where synthetic and original effects agree are candidates for held-out validation, not automatic wins. Disagreements may expose a weak simulator, a transcription error, a population shift, or a fragile original result. Human

replications adjudicate among those possibilities. In turn, their results improve both the evidence map and the evaluation set for future simulators.

6 From 154 studies to three disciplines

The current Ditto repository contains 154 runnable transcriptions of published conjoint and choice experiments. They span multiple applied domains and include encoded designs, human estimates, provenance, and machinery for execution and scoring. This is the current seed. It is not a representative sample of economics, sociology, or psychology; it does not include every experimental design; and it does not establish a corpus-wide fidelity result.

Growth should proceed by widening the design envelope only after the preceding one is auditable:

1. **Conjoint and choice.** Freeze inclusion rules, version the 154-study seed, publish complete and incomplete records, and report model-by-study results including failures.
2. **Randomized behavioral experiments.** Add survey, laboratory, online, and field experiments with explicit treatment arms and recoverable outcomes.
3. **Quasi-experimental evidence.** Add designs such as regression discontinuity, difference-in-differences, and instrumental variables only with their identifying assumptions and diagnostics intact.
4. **Continuous Atlas.** Ingest new publications and replications, maintain study families, and expose a queryable graph of treatment–outcome edges by population and context.

Coverage is measured, not announced. Each release reports the search universe, screened count, eligible count, executable count, synthetic-run count, human-replication count, and reasons for exclusion. "Every study" becomes credible only when a reader can see the denominator and the gaps.

7 Governance is part of the method

A system that compresses a literature can amplify its biases. The Atlas therefore needs governance rules as concrete as its software contract.

Provenance before plausibility. An unsupported value remains missing. Curator decisions are attributed and reviewable. Original authors can propose corrections, but no source controls the interpretation of a failed replication.

Failures stay visible. Aborted runs, low-information responses, design mismatches, null effects, and negative replications remain in release denominators. A map that displays only recovered effects is marketing, not evidence.

Identity stays contextual. A variable is defined by concept, bearer, experimental role, and measurement context. Respondent age, candidate age, and age used as a subgroup moderator are different variables even when a table calls all three "age."

Model versions expire. Synthetic results are immutable records tied to exact model and protocol versions. New models add evidence; they do not overwrite old misses.

Human subjects remain protected. The Atlas does not create permission to rerun harmful, deceptive, discriminatory, or privacy-invasive interventions. Ethical review, consent, data rights, and community constraints remain study-specific gates.

Standards remain plural. Sign agreement, effect calibration, predictive fit, and human replicability answer different questions. The interface must not collapse them into a single badge.

Open-science standards have shown how disclosure, registration, data sharing, and common reporting rules can make claims easier to inspect (Nosek et al., 2015). The Atlas extends that logic from publication artifacts to executable causal records.

8 The outcome

The outcome is not a paper count or a larger synthetic focus group. It is a public instrument for answering four questions about a causal claim:

1. What exactly was changed, measured, and estimated?
2. Can the study be executed from the available record?
3. Does a specified synthetic population recover the human effect, with uncertainty?
4. Has the effect itself survived an independent human replication, and for whom?

Today those answers require bespoke literature review and reconstruction. The Causal Atlas would make them properties of a versioned object. Researchers could find adjacent effects and unresolved contradictions. Model builders could evaluate synthetic populations on held-out interventions rather than plausible behavior. Funders could direct human replication toward consequential claims and known simulator failures. Decision makers could inspect the evidence behind an estimate before acting on it.

The ambition is deliberately larger than the current corpus: replicate every eligible study in economics, sociology, and psychology, then keep replicating as the literature and the world change. The discipline is equally important: start with 154, label it as 154, publish the misses, and never confuse a synthetic rerun with human ground truth. If those rules hold, replication stops being a sequence of heroic projects and becomes durable infrastructure for cumulative social science.

References

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. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Marcel Binz, Elif Akata, Matthias Bethge, et al. A foundation model to predict and capture human cognition. *Nature*, 644:1002–1009, 2025. doi: 10.1038/s41586-025-09215-4.

Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351 (6280):1433–1436, 2016. doi: 10.1126/science.aaf0918.

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2:637–644, 2018. doi: 10.1038/s41562-018-0399-z.

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. doi: 10.1093/pan/mpt024.

Brian A. Nosek, George Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck, Christopher D. Chambers, Gilbert Chin, Garret Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015. doi: 10.1126/science.aab2374.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716.

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

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

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. doi: 10.48550/arXiv.2408.16073.