PUBLICATIONS · INSTRUMENT AND FINDINGS

THE AGENTIC BEHAVIOR OBSERVATORY

READING WHAT A REPOSITORY MODELS, AND WHAT IT LEAVES OUT

Ethical Tech CoLab
August 2026

Analysis and write-up by the Ethical Tech CoLab. Figures computed from the committed corpus of 22 repositories, read on 26 August 2026.

> A repository that simulates a population makes a claim about who that population contains, and the claim is legible in its source. This report describes an instrument that reads it, and what the instrument found across 22 repositories: work that models age and disability far more often than income or literacy, an evaluation layer that is strongest exactly where the subject matter is weakest, and eight repositories whose behavioral findings rest on a language model they never name.

KEY FIGURES

22

repositories analyzed, 15 of them the CoLab's own and 7 reference frameworks

62

signals across six axes, each traceable to the file and line it fired on

3 OF 22

repositories that model income; 19 model children

8

repositories doing model-driven behavioral work that name no model version

01

THE QUESTION THE INSTRUMENT EXISTS TO ANSWER

When a system generates or evaluates a synthetic population at scale, it is making a claim about who that population contains. The claim is rarely written down. It is implied by which attributes the code carries, which behaviors the agents are allowed to have, and which differences between people the model treats as differences at all. A population dimension nobody models is a population nobody simulates, and an unmodeled dimension is an implicit assertion that it does not matter to the outcome.

That claim is legible in the source. The Agentic Behavior Observatory reads it. Paste a GitHub repository URL and it returns an evidence-linked account of how that repository models agentic behavior, scored on six axes, with the demographic dimensions and the model versions it rests on pulled out. Every point of every score links to the file and the line where its signal fired.

What a score is not. A score is signal coverage, not a quality judgment. A small, sharp repository can and should score lower than a sprawling framework, and the tool says so on its own front page. The number answers one question only: how much of the taxonomy's vocabulary does this repository exhibit. Whether the work is good is a question for a reader, not a regular expression.

02

HOW THE READING IS DONE

The analyzer fetches a repository's metadata and file tree from the GitHub API, then reads up to 120 file bodies from the raw content host, chosen by a heuristic that favors manifests, prose, and modeling-core filenames. Text source in roughly 25 extensions counts, including the single-file HTML applications much of the CoLab's work ships as. Generated bundles and lockfiles are skipped. Nothing is cloned and nothing is executed.

Each file is matched against 62 signals grouped into six axes. A signal is a dependency declaration, a file path, or a pattern in source and prose, and it carries a weight and a human-readable label. An axis score is the share of that axis's weighted signal set the repository covers, expressed on a 0 to 100 scale.

AXIS	SIGNALS	WHAT IT DETECTS
Agent-based simulation	16	Mesa, AgentPy, PettingZoo, NetLogo, Repast, SimPy; agent classes, schedulers, step loops, spatial environments
Synthetic data generation	10	SDV, CTGAN, synthcity, Faker, Gretel; population synthesis and IPF, census seeds, differential privacy, generation at scale
Model-based behavioral modeling	10	Anthropic and OpenAI SDKs, LangGraph, AutoGen, CrewAI, DSPy, local runtimes; personas, memory streams, generative-agent architectures, silicon sampling
Reinforcement learning	8	Gymnasium, Stable-Baselines3, RLlib, TorchRL; named algorithms, reward machinery, RLHF and DPO, the reset and step contract
Evaluation and validation	8	Tests, fidelity metrics such as KS, Wasserstein and TSTR, sensitivity analysis and ablations, seeded runs, experiment tracking, bias and representativeness audits
Context isolation	10	Private and privileged instructions, BATNAs and red lines; visibility and disclosure rules; per-agent scoped context; fresh sessions and state resets; independent replications; staying in character and knowledge-cutoff handling; contamination and leakage checks; blind evaluation

The six axes and the signals behind each. Counts are of distinct signals, not of matches.

The headline relevance score uses only the three subject axes, because the other three describe how work is done rather than what it is about.

```
relevance = 0.6 x max(subject axes) + 0.4 x mean(subject axes)
```

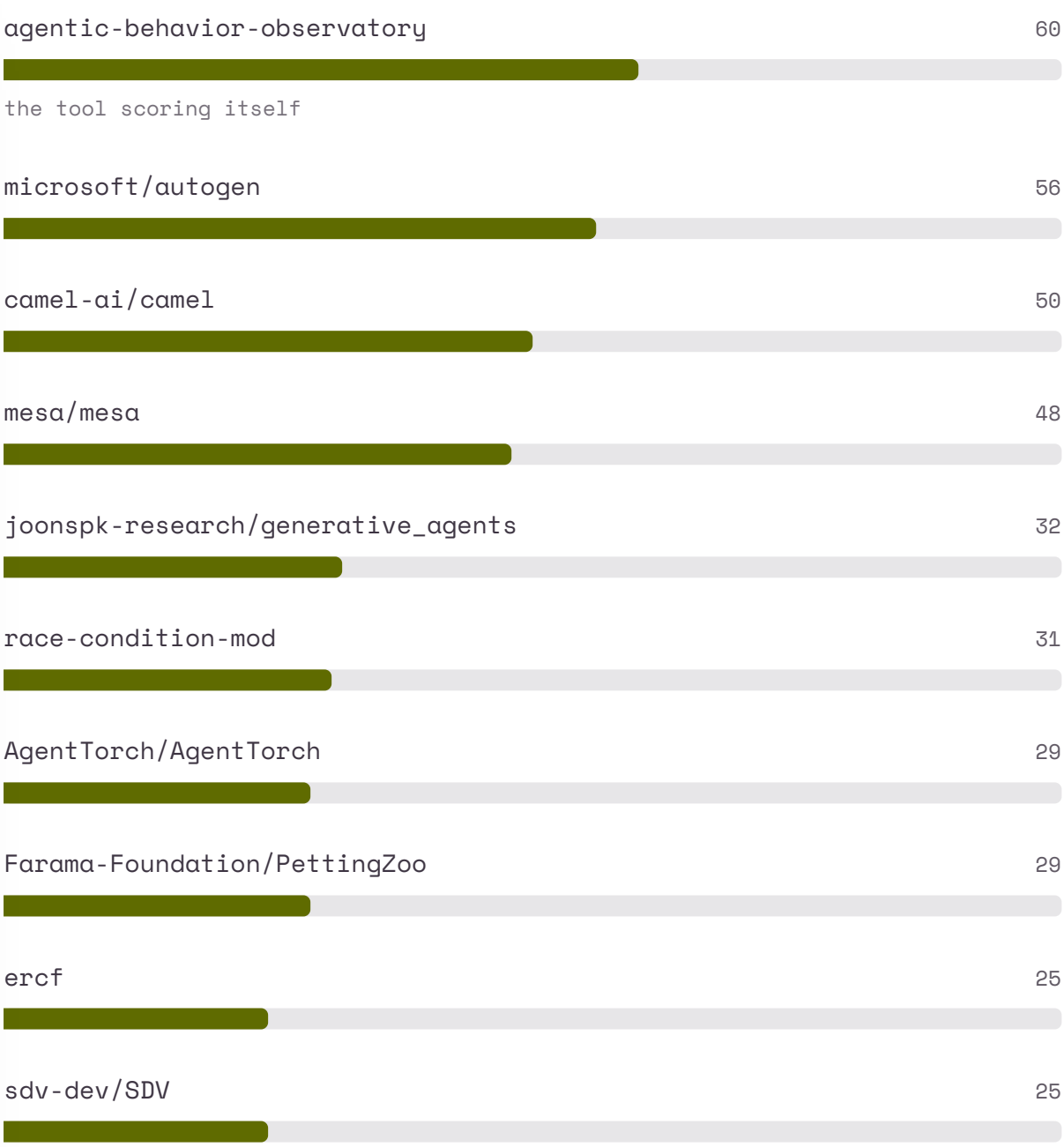
Subject axes are agent-based simulation, synthetic data generation, and model-based behavioral modeling. The max term keeps a repository that does one of them thoroughly in scope; the mean term rewards work that spans them.

One taxonomy, two runtimes. The scoring logic exists once, in signals.py, and is exported to the browser as JSON. The command-line analyzer that writes the committed corpus and the in-browser analyzer that scores a repository you paste therefore score identically. A visitor can reproduce any published number without installing anything.

03

THE CORPUS

The committed corpus is 22 repositories: 7 reference frameworks that define the field's vocabulary, and 15 Ethical Tech CoLab projects, mostly evacuation and negotiation simulators, plus the CoLab's agent work and its evaluation tooling. The observatory analyzes itself as well, which is discussed as a limitation below rather than claimed as a validation.



Relevance across the corpus. The reference frameworks cluster at the top not because they are better work, but because they contain more of the taxonomy's vocabulary, which is what the score measures.

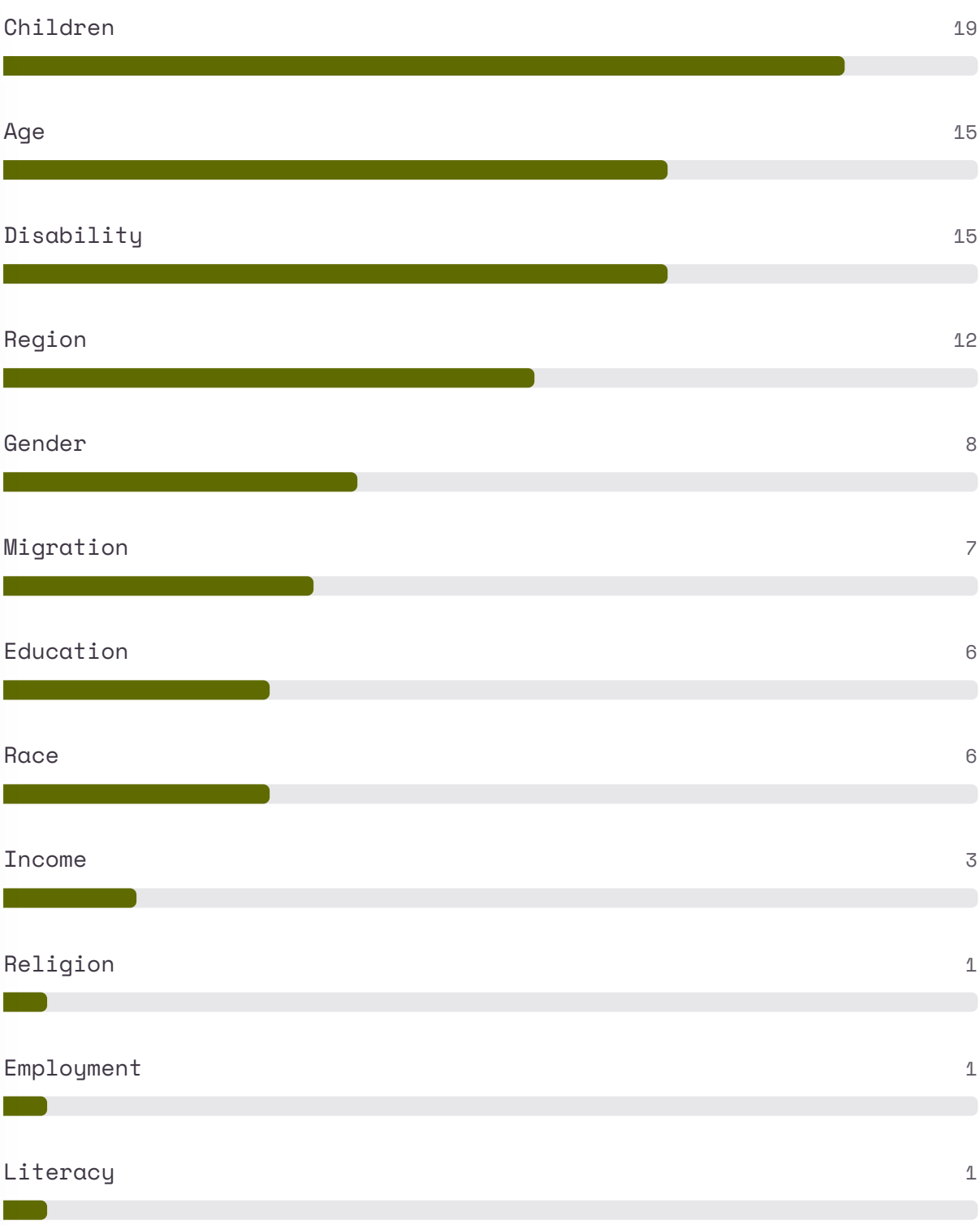
DATA TABLE ▸

Eleven of the 22 have model-based behavioral modeling as their strongest subject axis, 8 agent-based simulation, and 3 synthetic data generation. That distribution is a fact about which repositories were chosen, not about the field, and the corpus is small enough that a single addition moves it.

04

WHAT THE CORPUS TURNED OUT TO MODEL

Across all 22 repositories the analyzer found 33 distinct demographic dimensions in code and prose. They are not evenly distributed, and the shape of the imbalance is the most substantive finding in this report.



Repositories mentioning each demographic dimension, out of 22. Visible bodily and household attributes dominate; economic position is close to absent.

DATA TABLE ▸

Populations with bodies but no economics. Nineteen of 22 repositories model children and 15 model disability, both of which change how a person moves in an evacuation. Three model income, one models employment, and one models literacy. In a body of work substantially about who gets out of a disaster and who does not, the attributes that determine whether a household has a car, can afford to leave early, or can read the warning are the ones least often represented. The instrument cannot say whether that is an oversight or a defensible modeling choice. It can say the choice is being made silently in most of these repositories.

Eight repositories name no model. The analyzer extracts every model identifier it finds. Across the corpus it found 42 distinct versions, led by gpt-4o-mini in 7 repositories. Eight repositories doing model-driven behavioral work name no version anywhere the analyzer could read. Behavioral findings drift between model versions, so a result whose model is unrecorded cannot be reproduced later even by the people who produced it.

The corpus is better at isolating contexts than at modelling economies. Context isolation was added to the taxonomy after the findings above, to ask whether a repository has any vocabulary for keeping information where it belongs: one agent out of another's private brief, one run out of the next, the model's own training out of the persona it was given, the test item out of the prompt meant to test it. Median across the corpus is 40. The clearest case is the Diplomatic Simulator at 73, its highest axis by a wide margin and not an accident of vocabulary: every delegation holds private instructions, a BATNA and red lines, and the table carries explicit rules for what crosses between parties. Its own paper states the principle outright. At the other end the Stanford generative agents repository scores 0, which is worth sitting with, because the memory stream is the architecture and nothing in it names a boundary between one agent's stream and another's.

Evaluation is strongest where the subject is weakest. Median evaluation across the corpus is 48, higher than any subject axis: the medians for agent-based simulation, synthetic data generation, and model-based behavioral modeling are 11.5, 7 and 17. AgentTorch scores 84 on evaluation against 29 relevance. The pattern is partly an artifact of the taxonomy, since test files and CI configuration are easy vocabulary to detect, and partly real: it is easier to add a test suite than to model an economy.

The landmark exception. The Stanford generative agents repository, the most cited artifact in this literature, scores 0 on evaluation. It is research code released to accompany a paper rather than a maintained framework, and

the paper carried the validation. The score is accurate and would be misread as a verdict by anyone who took it for one, which is the clearest illustration in the corpus of why coverage is not merit.

05

LIMITS WORTH STATING

The instrument is a vocabulary detector. Reading it as anything more would reproduce exactly the error it was built to expose.

- Regular expressions match vocabulary, not meaning. A repository that discusses differential privacy without implementing it fires that signal. Every signal therefore links to its evidence, and the evidence, not the score, is the finding.
- The word party is counted as a demographic dimension in 17 repositories, which is the corpus's clearest false positive: in a negotiation simulator a party is a side at the table, not a political affiliation. It is left in rather than special-cased, because the general problem it illustrates does not go away by patching one term.
- The demographic vocabulary is a fixed English list of 39 terms, so it under-reports populations described in other words or other languages. That limit is itself a finding about who the tooling was built by.
- Up to 120 files are read per repository. Large repositories are sampled, not read whole, and every report carries files_read, files_eligible and files_total so the sampling is never silent.
- The observatory scores itself 60, first in the corpus, and rose from 54 when the context-isolation signals were added to the very file it analyses. A repository whose contents are the taxonomy will match the taxonomy, so this number is close to tautological and should not be read as the tool validating itself. It is the sharpest available demonstration of the limit above it.
- Twenty-two repositories is a small corpus, and 15 of them come from one organization. The dimension counts above describe this corpus. They are a prompt to check your own, not a measurement of the field.

06

USING IT, AND EXTENDING IT

The dashboard runs the analysis in the browser. Two calls to the GitHub API for metadata and the file tree, then file bodies from the raw content host, which is CORS-open and not rate limited: roughly thirty repositories an hour with no token at all. Results stay in the browser and are marked live. Downloading the JSON is how one joins the shared corpus.

The command-line analyzer writes the committed corpus. It has no dependencies beyond the Python standard library, clones nothing, and picks up a GitHub token from the environment or from the gh CLI if one is available.

Teaching it a framework. The taxonomy is one file. Adding a simulation framework is one tuple naming its axis, key, kind, pattern, weight and label. The build step exports the same taxonomy to the browser, so a single edit updates both runtimes and no score can differ between them.

Analyzing the CoLab's own corpus exposed three defects in the taxonomy, all since fixed: HTML and JSX source was not being read at all, the file size cap excluded single-file applications, and the agent-detection patterns assumed agents are called agents rather than members, families, or residents. Any repository analyzed before those fixes scored misleadingly low, which is the argument for versioning a taxonomy the way one versions a model.

References

SOURCES

01 Ethical Tech CoLab. Agentic Behavior Observatory: dashboard, methodology, and committed corpus.

02 Ethical Tech CoLab. agentic-behavior-observatory: analyzer, taxonomy (analyzer/signals.py), and per-repository reports (data/reports/).

03 Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative Agents: Interactive Simulacra of Human Behavior.

04 Mesa. Mesa: agent-based modeling in Python.

05 Microsoft. AutoGen: a programming framework for agentic AI.

06 Farama Foundation. PettingZoo: an API for multi-agent reinforcement learning.

07 DataCebo. SDV: the Synthetic Data Vault.

AgentTorch. AgentTorch: differentiable large-scale agent-based models.

CAMEL-AI. CAMEL: communicative agents for mind exploration of large language model society.